Applied bibliometrics. From data to publication

English Edition

AG
EDITOR

Applied bibliometrics. From data to publication (English Edition)
First Edition

Annier Jesús Fajardo Quesada

Eduardo Antonio Hernández González

René Herrero Pacheco

Lisannia Virgen Beritán Yero

AG
EDITOR

Copyright Page

This AG Editor imprint is published by AG Editor.

The registered company is **AG Editor SAS, Montevideo, Uruguay.**
For more information, see AG Editor's Open Access Policy: https://www.ageditor.org/editorial-policies.php

ISBN (English edition, eBook): 978-9915-9680-7-0

ISBN (Spanish edition, eBook): 978-9915-9680-6-3

This title is available in both print (softcover) and digital (open access PDF) formats.

DOI: 10.62486/ 978-9915-9680-7-0

Publisher: Javier González Argote
Chief Executive Officer: editorial@ageditor.org
Editorial Director: Emanuel Maldonado
Editorial Coordinators: William Castillo González; Karina Maldonado
Production Manager: Adrian Vitón Castillo

Legal Deposit: National Library of Uruguay — Law No. 13.835/1970 and Decree No. 694/1971
ISBN Record: National ISBN Agency (Uruguay) — Filing No. 58241

Cataloging Data

Cataloging-in-Publication Data (CIP):
AG Editor
Applied bibliometrics. From data to publication (Spanish Edition) / Annier Jesús Fajardo Quesada, Eduardo Antonio Hernández González, René Herrero Pacheco, Lisannia Virgen Beritán Yero. – Montevideo: AG Editor, 2025.

First Edition.

ISBN (English edition, eBook): 978-9915-9680-7-0

ISBN (Spanish edition, eBook): 978-9915-9680-6-3

THEMA classification codes:
GLF — Information technology, Internet and electronic resources in libraries

Editorial Notice and Acknowledgments

The publication of this book is part of AG Editor's mission to promote open, ethical, and rigorous scientific communication across all areas of knowledge.

All books published by AG Editor undergo a double-blind peer-review process and a technical evaluation in accordance with the publisher's editorial policies, aligned with COPE and ICMJE standards.

AG Editor acknowledges the valuable collaboration of authors, reviewers, designers, and production teams who made this publication possible.

“La verdad es la verdad, dígala Agamenón o su porquero”

El Porquero de Agamenón

PROLOGUE / PRÓLOGO

Bibliometrics as a Pilot in the Storm of Open Science.

We live in an era of information explosion. Every 6 seconds, a new scientific article is published somewhere in the world. In 2024 alone, Scopus recorded more than 5 million indexed studies. Faced with this deluge of knowledge, researchers, institutions, and governments face a critical challenge: how to navigate this ocean without sinking into irrelevance or redundancy?

Bibliometrics emerges not as an academic luxury, but as a tool for intellectual survival. This discipline, which quantitatively analyzes the production and dissemination of knowledge, has become the essential radar for:

- Prioritizing resources in universities with dwindling budgets
- Detecting scientific fraud in fields such as medicine

But today its role is even more vital. Open science, with its mandate of transparency, open access, and reproducibility, has made bibliometrics the guardian of academic integrity. When the Leiden Manifesto (2015) and the DORA Declaration (2012) demand that research be evaluated for its real impact, not for the impact factor of its journals, it is responsible metrics that enable this promise to be fulfilled.

This book was created to empower you in this paradigm shift. Not as a distant theorist, but as a practical guide that transforms data into decisions. Here you will not find abstract formulas, but proven protocols for:

- Identifying emerging lines of research before competitors
- Demonstrating the social impact of your work beyond citations
- Avoiding biases that distort university rankings

This work was crafted for two profiles of knowledge explorers:

1. Researchers

For those of you who write articles in the early hours of the morning:

How can you demonstrate that your work on nanotechnology applied to crops deserves funding?

How can you identify key collaborators in your country for your project?

Here you will learn to:

- Use VOSviewer to map co-authorship networks
- Calculate your discipline-corrected h-index
- Document your impact with alternative metrics (e.g., derived public policies)

2. Students

For those of you starting your academic journey:

How can you choose a supervisor with a proven track record (not just fame)?

Which journals should you prioritize for your first publication?

You will find:

- Tutorials in Google Colab for analysis without installing software
- Step-by-step instructions with Bibliometrix
- Alerts about bibliometric predators (journals that inflate metrics)

From the Machete to the Satellite

This book is a methodological expedition designed in five progressive stages

Part I: Fundamentals of Navigation

Chapter 1 equips you with conceptual GPS: differences between bibliometrics, scientometrics, and altmetrics.

Chapter 2 is your historical compass: from Garfield's citation indexes (1955) to the open metadata revolution.

Part II: Preparing for the Journey

Chapter 3 is your digital toolbox: installing Python/R even on slow computers.

Chapter 4 is your survival kit: ethical data retrieval in Scopus, WoS, and PubMed.

Part III: Exploring Complex Territories

Chapter 5 unfolds your topographic map: critical calculation of 15 bibliometric indices (with replicable codes).

Chapters 6, 7, and 8 are your mapping drone: creating and interpreting graphs with VOSviewer and CiteSpace.

Chapter 9 integrates the rest of the indicators to enhance your results.

Part IV: Communicating Findings

Chapters 10 and 11 are your travel log: how to write methods, results, and discussions that pass rigorous reviews.

Part V: Future, Resources, and Applications

Chapters 12 and 13 project your telescope toward generative AI and the web.

Appendices are your emergency kit: glossary and sites of interest.

The Commitment: Rigor without Dogma

This manual rejects two dangerous extremes: metric fetishism that reduces science to numbers and anti-quantitative contempt that judges measuring impact as "dehumanizing."

Instead, we offer you a three-dimensional approach:

- Technical (Python/R codes updated to 2025)
- Critical (alerts on gender, language, and geography biases)
- Ethical (aligned with open science and responsible evaluation)

At the end of this journey, you will not be just a passive user of metrics. You will be a navigator capable of:

- Creating valid bibliometric studies
- Interpreting complex graphs with a clinical eye
- Debating scientific policies with hard evidence

21st-century science demands more than data producers: it requires interpreters of meaning. This is your compass to stay on course.

Annier Jesús Fajardo Quesada
Havana, October 2025.

INDEX / ÍNDICE

Part I / Parte I

INTRODUCTION AND FOUNDATIONS

INTRODUCCIÓN Y FUNDAMENTOS

AG
EDITOR

Chapter 1 / Capítulo 1

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch01

Introduction to Bibliometrics / Introducción a la Bibliometría

1.1. Definition and Objectives of Bibliometrics

The Science that Measures Knowledge

Bibliometrics is a scientific discipline that applies quantitative methods to systematically analyze the production, dissemination, and impact of academic knowledge. The term, which derives from the Greek words 'biblion' (book) and 'metron' (measure), has evolved to encompass the study of a wide range of contemporary academic products, from scientific articles and patents to doctoral theses and data sets. This methodological approach represents a robust alternative to traditional bibliographic reviews, distinguished by three fundamental attributes that ensure its scientific rigor.[1]

Methodological objectivity is based on the analysis of concrete empirical evidence, such as citation patterns, collaboration networks, and standardized impact indicators. This feature allows us to overcome the limitations of purely qualitative assessments, relying on verifiable data and standardized metrics. Operational scalability enables processing massive volumes of scientific literature with specialized computational tools, analyzing thousands of documents in a short time. Finally, analytical reproducibility is ensured by standardized protocols, such as the PRISMA guidelines for systematic study selection, which enable independent verification of results.[1,2]

1.1.1. Objectives: Beyond Counting Citations

Bibliometrics is not limited to measuring citations; it is an analytical discipline that seeks to understand, evaluate, and optimize scientific production. Its five fundamental objectives enable researchers, institutions, and scientific policymakers to make evidence-based decisions.[3]

Evaluating scientific impact

The primary objective of evaluating scientific impact has evolved significantly, transcending the reductionist view that equated it exclusively with the counting of bibliographic citations. While this traditional indicator remains an essential element, offering tangible evidence of the specialized community's assimilation of research and its contribution to advancing disciplinary knowledge, the contemporary conception of impact is notably richer and more nuanced.

Modern bibliometrics has come to recognize that genuine scientific impact is not a one-dimensional phenomenon but a ripple effect that manifests in multiple ways. Therefore, its evaluation now integrates a broader spectrum of metrics that capture the influence of research along various axes. On the one hand, altmetrics act as a thermometer of practical utility and knowledge transfer beyond the academic ivory tower. An article that informs a clinical guideline or an environmental regulation has a profound scientific impact, even if its number of citations is moderate.

On the other hand, tools such as Dimensions have refined this analysis by enabling the synergistic integration of traditional citations with data from patents, funding projects, and datasets. This integration not only enriches the evaluation, but also allows the research life cycle to be traced: from the original idea, through its development (funded by projects), to its protection (patents) and, finally, its dissemination and influence (citations and use in new research). In this way, it is possible to discern whether a work, rather than being cited, has served as a cornerstone for a new field of technological development or a fruitful line of research.

In short, the current goal of evaluating scientific impact is no longer satisfied with counting echoes in the literature, but with understanding the nature and scope of the contribution. It is about differentiating between mere visibility and substantive influence that accelerate discovery, inform application, and ultimately consolidate the legacy of research in the edifice of human knowledge.

Mapping knowledge structures

Beyond measuring impact, bibliometrics has a unique ability to map the vast and dynamic territory of scientific knowledge. Academic production is not an isolated set of papers, but an organic ecosystem in constant evolution, where ideas, disciplines, and researchers interconnect to form complex structures. The goal here is to reveal these invisible architectures, identifying the intellectual currents that are gaining strength, those that are waning, and the empty spaces where innovation can flourish.

Techniques such as co-word analysis, which identifies concepts that frequently appear together, and bibliographic coupling, which links documents that share standard references, provide powerful lenses for observing this landscape. These methodologies are not limited to listing topics; they detect patterns, reveal the anatomy of disciplines, and show how fields of study emerge, converge, or fragment.

The true power of this analysis is realized in visualization tools such as VOSviewer or CiteSpace, which will be discussed in later chapters. These programs transform massive publication data into intuitive visual maps, where thematic clusters are grouped into colored clouds, and the proximity between nodes indicates the strength of their relationship. A glance at these maps can reveal emerging trends before they consolidate, point to fruitful interdisciplinary collaborations, or, conversely, highlight saturated areas of research where it is difficult to make an original contribution.

This mapping is not merely a descriptive exercise; it is a strategic tool of the first order. For the individual researcher, it serves as a compass pointing to “niches of opportunity” or gaps in the literature, thereby guiding the formulation of truly novel research questions. For institutions and funding agencies, these maps provide an objective basis for designing science policy, allocating resources intelligently, and fostering collaborations that strengthen research ecosystems with greater potential.

Optimizing academic resources

In a context of finite resources and growing competition, science management cannot rely solely on intuition or historical inertia. Bibliometrics emerges here as a fundamental tool for strategic management, bringing the rigor of data to critical decisions about funding, hiring, and institutional planning. Its goal is to ensure that investment in research, whether public or private, is directed where it can generate maximum scientific and social returns.

Bibliometric analysis enables the identification of areas of knowledge with exceptional dynamism, promising growth rates, or tangible societal transfer potential. In this way, funding agencies, such as the National Science Foundation (NSF), can prioritize grants in emerging fields with a greater likelihood of future impact. At the same time, this data helps avoid the dispersion of efforts in saturated or narrow fields of research. A revealing fact that underscores the need for efficiency is that, according to a PLOS ONE (2024) study, approximately 60 % of scientific articles are never cited. This figure invites deep reflection on academic productivity and questions the efficiency of some lines of research, promoting a more qualitative and strategic approach.[4]

Beyond subject areas, bibliometrics allows research capital to be evaluated from a multidimensional perspective. Institutions no longer focus solely on a candidate's publication volume. They analyze metrics of influence, such as citation counts, and, crucially, map their collaboration networks. Identifying researchers with strong and established international networks has become a key factor, as this collaborative science often translates into more ambitious, innovative, and visible projects.

In short, bibliometrics has transformed scientific governance. It provides universities and research centers with an evidence-based "dashboard" for making informed decisions: from hiring the talent that best enhances their research ecosystem to reorienting their departments toward niches of excellence. In essence, it is about replacing conjecture with analysis, optimizing each resource to strengthen the R&D&I system as a whole.

Detecting irregularities

At its most critical, bibliometrics transcends its evaluative function to become an essential instrument of surveillance and quality control within the scientific ecosystem. Faced with growing pressure to publish and the perverse incentives that can arise from an overly quantitative system, there is a temptation to manipulate performance indicators. This is where bibliometric techniques act as a "scientific police force," detecting and deterring practices that compromise academic integrity.

Among the most common irregularities are excessive self-citation, where an author or group disproportionately cites their own work to inflate its impact artificially; publication in predatory journals, characterized by opaque review processes and anomalous citation patterns; and the formation of citation cartels, networks of journals or authors who agree to cite each other to inflate their metrics artificially.(5)

To combat these distortions, sophisticated algorithmic tools have been developed. Platforms such as Scimago Journal Rank monitor journal behavior, flagging those with suspicious citation profiles, such as an abnormal and sudden increase in their impact factor. More advanced algorithms developed in environments such as Python, including the Citation Cartels Detector, can analyze large volumes of data to identify clusters of publications exhibiting patterns of circular or mutually beneficial citations that do not reflect genuine scientific merit.(5)

The rigorous application of these analyses is vital to protecting fairness in scientific evaluation. It allows funding agencies, promotion committees, and journal editors to make decisions based on refined data, rewarding real merit and isolating fraudulent practices. In this way, bibliometrics not only measures excellence but also actively contributes to defending it, ensuring that resources and recognition are allocated fairly and transparently.

Promoting open science

The emergence of open access (OA) has radically reshaped the foundations of scientific communication, democratizing knowledge by removing paywalls for readers. In this context, bibliometrics has established itself as the indispensable discipline for quantifying the true scope of this transformation, offering crucial insights beyond simple publication counting.

When analyzing the geographical distribution of gold OA (where articles are published immediately and free of charge in the journal), the data reveal clear leadership from European countries such as the United Kingdom, Germany, and the Netherlands, driven by strong institutional and financial mandates. They are closely followed by nations such as Brazil and

Indonesia, which have promoted powerful national repositories. This bibliometric map not only shows adoption, but also a clear geo-economic divide in the ability to bear publication costs (APCs), posing a crucial challenge for global equity.

The impact of OA is particularly significant in disciplines such as the humanities. Traditionally confined to a slower communication cycle and lower citation rates, open access has dramatically boosted its visibility. By removing the paywall, humanities research reaches a much wider audience: educators, cultural professionals, journalists, and interested citizens. Modern bibliometrics, through altmetrics, captures this expanded influence, demonstrating that impact in these fields is measured not only by citations but by their ability to permeate public debate and culture.

The so-called "OA citation advantage," backed by studies indicating that open access articles can receive up to 50 % more citations, for example, is a powerful argument for its implementation. However, this advantage is not uniform. Its magnitude varies substantially across disciplines and regions, underscoring the need for differentiated policies.

To navigate this complex landscape, tools such as Unpaywall and the European Union's Open Access Panel have become essential. These platforms track the status of millions of articles in real time, allowing trends to be visualized, barriers to be identified, and the effectiveness of national policies to be evaluated. Thus, bibliometrics provides the evidence needed to design strategies that not only promote open access but also do so intelligently and equitably, closing gaps and ensuring that knowledge is a common good, not a privilege.[6]

1.1.2. Ethical limits: What bibliometrics should NOT be

Bibliometrics is a powerful tool, but its uncritical use can distort scientific evaluation. Recognizing its ethical limits is essential to prevent quantitative indicators from becoming ends in themselves, perpetuating inequalities, or replacing qualitative analysis. Below are three key principles that every researcher or institution should consider when applying bibliometric metrics.

It is not an end in itself: metrics should serve knowledge, not replace expert judgment

Metrics are means, not ends. Reducing the value of research to its number of citations or h-index ignores key dimensions such as its social relevance, originality, or methodological rigor. For example, theoretical studies in philosophy may have less immediate impact than applied work in engineering, but their long-term influence can be profound. Academic evaluation should combine quantitative indicators with qualitative peer reviews. Cases such as the *San Francisco Declaration on Research Assessment* (DORA) warn against the simplistic use of metrics in hiring or funding decisions, promoting holistic evaluation instead.

It is not neutral: it reflects gender, language, and geographical biases

Bibliometric data are not objective; they reproduce the structural inequalities of science. For example:

- Gender bias: only 33 % of authors cited in artificial intelligence are women (*Nature* study, 2023), which renders key contributions invisible.[7]
- Language bias: 90 % of publications in Scopus are in English, marginalizing research in other languages, especially in the humanities and social sciences.[8]
- Geographic bias: databases favor journals from Europe and the US, underrepresenting Africa, Latin America, and Asia, even in areas such as biodiversity and tropical medicine.[9]

Ignoring these biases perpetuates cycles of exclusion. Solutions such as Cite Black Women or the Latin American Citation Index seek to correct these imbalances, but a conscious effort is required to diversify the sources of analysis.

It is not foolproof

Ctation errors and manipulation distort results. Bibliometric metrics are not perfect and can be contaminated by questionable practices or systemic flaws such as incorrect citations, inflated self-citations, and predatory journals.

Tools such as Citation Context Analysis or anti-manipulation algorithms help detect these problems, but transparency and human auditing remain indispensable.

This practical guide not only instructs on the calculation of bibliometric indicators but also fundamentally promotes their systematic questioning, ensuring that bibliometrics enriches scientific practice without simplifying or distorting it through reductionist applications. The balance between analytical use and critical awareness is the key to the ethical and practical implementation of these tools in the contemporary research ecosystem.

> *"Bibliometrics is like a thermometer: it measures the fever of science, but it does not diagnose the disease or prescribe the cure."*
> *Adapted from Loet Leydesdorff.*

Questions that Bibliometrics Answers

1. Who are the key players in my field? → Co-authorship analysis
2. Is my research aligned with global trends? → Thematic maps
3. How can I demonstrate my impact to non-expert evaluators? → h-index + altmetrics

1.2. Classification of bibliometrics

Bibliometrics, like any scientific discipline, requires divisions so as not to get lost in the immensity of data. These allow us to organize and make sense of the vast universe of publications and indicators. Far from being a mere abstract taxonomy, this organization is essential for applying the right tool to each research question, whether measuring the impact of a journal, the productivity of a group, or the soundness of a theory. Without this order, we would be faced with a chaos of numbers without narrative.

According to the object of evaluation

Bibliometrics is structured around different objects of analysis, the most prominent being scientific journals and researchers. In the first case, prestige indicators such as the Impact Factor (JCR) or the SCImago Journal Rank (SJR) are used, which measure the average influence of articles published in a journal and are frequently used for decision-making in editorial and evaluation policies. In the case of authors and their research groups, metrics such as the h-index or the i10-index are used to synthesize their productivity and the perceived impact of their work in a single figure, offering an overview of their career and relevance within the academic community. This distinction is crucial in order not to confuse the impact of a periodical publication with the individual merit of those who publish in it.[10]

According to the nature of the measure

A complementary perspective organizes indicators according to the quality they measure. Productivity metrics, such as the total number of publications, provide a basic quantification of research output but do not account for its scope or significance. A deeper level of analysis is provided by impact or visibility metrics, the most direct example of which is citation counts, which seek to reflect the usefulness and adoption of scientific findings by the academic community. Finally, collaboration metrics, often represented by co-authorship network maps, transcend the quantitative to examine patterns of intellectual cooperation, revealing the structure and dynamics of relationships between researchers, institutions, and countries.[11]

According to the temporal dimension

This text introduces temporality as a fundamental classification criterion, arguing that the time window analyzed substantially conditions the interpretation of any indicator. Time series analysis is the most robust approach, as it allows us to observe the evolution of indicators over multiple periods. As discussed in later chapters, bibliometric indicators can be combined over time to identify growth trends, turning points in a field of study, or the consolidation of a researcher's impact, providing a dynamic and contextualized narrative. In contrast, a short-term analysis, such as a single year, provides only a static snapshot. This fragmented view is unable to distinguish between a sustained trend and a one-off fluctuation. It is generally insufficient for assessing the strength of a line of research or a scientist's trajectory.

1.3. Key differences: bibliometrics vs. scientometrics vs. altmetrics

Three Lenses for Measuring Science

In the quantitative research ecosystem, **bibliometrics**, **scientometrics**, and **altmetrics** are sister disciplines with complementary approaches. While they all measure the impact of knowledge, they do so from different perspectives. This section unravels their fundamental differences with practical examples and typical applications.

1.3.1. Bibliometrics: the analysis of published works

Bibliometrics is the quantitative study of academic documents, such as articles, books, and patents, and their relationships through citations, co-authorship, and keywords, addressing all its fundamental elements throughout this book. It focuses mainly on formal publications in indexed journals and uses classic indicators, such as citation counts, the h-index, and the impact factor, to measure impact and visibility. Its practical applications include evaluating individual researcher productivity, identifying key journals in specific disciplines, and mapping scientific collaborations through the analysis of co-authorship networks, thereby offering objective tools for understanding and optimizing the dynamics of scientific production.

1.3.2. Scientometrics: The Science of Science

Scientometrics is the quantitative study of science as a social system, analyzing not only its documentary production, but also its patterns of growth, collaboration, and impact. Unlike bibliometrics, which focuses on publication and citation metrics, scientometrics takes a broader view, examining how scientific activity is structured, evolves, and is funded.

Primary focus: Knowledge networks and bibliometric laws

Scientometrics studies the connections between disciplines and the emergence of new areas of research using various methodological tools. Among these, the analysis of collaboration networks stands out, allowing us to visualize and understand how researchers, institutions, and countries interact in the generation of scientific knowledge. This approach reveals patterns of cooperation and information flows that shape the research landscape.

In addition, scientometrics relies on fundamental bibliometric laws to explain scientific phenomena. Lotka's Law describes the asymmetric distribution of academic productivity, demonstrating that a small group of researchers generates most of the scientific output. On the other hand, Bradford’s Law helps identify the limited core of journals that concentrate the most relevant publications within a specific field.

These theoretical and methodological principles enable us to uncover hidden dynamics in the scientific world, such as the concentration of knowledge in leading institutions and the emergence of new disciplines branching from traditional areas. Through this analytical approach, scientometrics provides a deeper understanding of the structure and evolution of scientific knowledge.

Practical applications: from trend prediction to R&D policies

Scientometrics has established itself as a strategic tool for decision-making in science policy, offering valuable insights into research dynamics that transcend traditional bibliometric analysis. A key aspect of its application is the early identification of emerging disciplines, where techniques such as co-citation analysis have demonstrated their predictive power. This was the case with artificial intelligence applied to medical diagnosis, where cross-disciplinary citation patterns anticipated its clinical adoption years before it appeared in the specialized literature.

In the field of science policy, scientometrics provides robust methodologies for evaluating the return on investment in R&D. Leading countries in innovation, such as South Korea, have implemented advanced scientometric systems that correlate scientific production indicators with economic metrics, allowing the impact of public subsidies in strategic areas such as nanotechnology to be quantified. These models have proven particularly useful for optimizing resource allocation in sectors with high technological potential.

Another field where scientometrics provides unique value is in the study of global scientific mobility. By analyzing publication patterns and institutional affiliations, it has been possible to map research talent flows with unprecedented accuracy. The data reveal complex phenomena, such as the Indian scientific diaspora, in which nearly 40 % of researchers who publish in collaboration with US institutions remain in the US, according to a recent study published in Nature (2023). These findings have profound implications for the design of talent retention policies and repatriation programs in developing countries.

Challenges and biases in scientometric analysis

Although scientometrics offers a powerful lens for understanding scientific enterprise, it is crucial to recognize that this lens is not perfectly transparent. Its analytical potential coexists with several structural and methodological limitations, which require cautious, critical interpretation of its results.

First, the discipline is entirely dependent on the quality and breadth of big data from global databases such as Scopus and Web of Science. These platforms, despite their comprehensiveness, are not neutral mirrors of global knowledge production. They systematically underrepresent research from developing regions, publications in languages other than English, and local journals, distorting the international map of science and perpetuating a geolinguistic bias.

Secondly, the sophistication of scientometric techniques is a double-edged sword. Complex mathematical models, such as network analysis to map collaborations or the application of Price’s Law to model the exponential growth of literature, require advanced statistical expertise. Incorrect application or a superficial interpretation of these models can lead to erroneous conclusions, creating a mirage of quantitative precision that hides simplifications or conceptual misunderstandings.

Finally, and more profoundly, scientometrics inherits and, at times, amplifies the biases inherent in the scientific system itself. The data it analyzes is not generated in an equitable vacuum; it reflects structural inequalities in funding, the dominance of English as a lingua franca, and academic power dynamics. Therefore, a collaboration map may show not only excellence but also exclusion, and an impact indicator may be measuring, in part, the preexisting visibility of an institution or country.

In conclusion, scientometrics is a formidable diagnostic tool, but not an oracle. Its actual value lies not in uncritical acceptance of its metrics, but in researchers, managers, and policymakers' ability to understand its assumptions, recognize its biases, and use its findings as an informed guide, never as a definitive verdict.

1.3.3. Altmetrics: Impact on the digital society

In a world where science is discussed on Twitter, applied in public policy, and goes viral in podcasts, traditional metrics such as the *Journal Impact Factor* or citation counts are no longer sufficient to capture the real influence of research. This is where **altmetrics** (*alternative metrics*) come in: they track the impact of studies in digital and non-academic spaces, offering a broader view of their social relevance.

Primary focus: science in the public sphere

Altmetrics addresses fundamental questions about the social impact of scientific research beyond the traditional academic sphere. This discipline analyzes how scientific knowledge is used and discussed across various public and digital spaces, offering a broader view of the real reach of research.

Mentions on social networks such as Twitter, Facebook, or Reddit reflect the level of interaction and public discussion that a piece of research generates. When scientific findings appear on Wikipedia or are discussed in specialized blogs, this demonstrates their penetration into general culture and their influence on the dissemination of knowledge. More significantly, references in government documents or reports from non-governmental organizations indicate the concrete impact of research on policy formulation and decision-making.

To measure these impacts, altmetrics uses specialized platforms such as PlumX, Altmetric.com, and Dimensions, which collect and analyze various types of data. These range from downloads on academic platforms such as ResearchGate and Academia.edu, which show the direct interest of the research community, to coverage in mass media such as the BBC or The New York Times, which indicates the public relevance of the work. In addition, the use of research in online courses on platforms such as Coursera or edX demonstrates its incorporation into formal and informal educational processes.

These altmetric indicators complement traditional citation metrics, offering a multidimensional view of scientific impact that ranges from academia to society at large, including politics, education, and the media.

Practical applications: from public policy to dissemination

Altmetrics have become an essential tool for assessing the real impact of scientific research in various fields. One of its central values lies in its ability to measure the immediate social impact of studies on urgent issues such as climate change, public health, or gender equality, which often generate intense debate on digital platforms and in the media long before they accumulate traditional academic citations.

This discipline is particularly valuable for quantifying the effectiveness of scientific outreach strategies, allowing the reach of initiatives such as social media campaigns to be measured. When a university shares a discovery via TikTok, altmetric metrics such as shares, likes, and comments provide concrete data on its societal penetration, information that academic citations could not offer in the early stages.

In addition, altmetrics play a crucial role in justifying funding for research projects. The presence of a scientific article in public policy documents, such as a state renewable energy plan, or its use by non-governmental organizations, tangibly demonstrates the practical usefulness of research. This type of evidence is essential for convincing governments and funding agencies of the applied value of scientific work, especially in areas where social impact is as vital as academic impact.

The flip side of the coin: the intrinsic challenges of altmetrics

The promise of altmetrics to capture social impact in real time is undoubtedly revolutionary. However, this dynamism poses a number of fundamental challenges that must be recognized to avoid a naive interpretation of its data.

One of the most significant risks lies in the quality and authenticity of the sources. Viral popularity is not synonymous with rigor. A tweet can accumulate thousands of likes due to a sensationalist headline, and bot campaigns or highly polarized debates can artificially inflate mentions, creating a mirage of relevance unsupported by academic or serious discussion. Distinguishing between noise and signal, between manipulation and genuine engagement, becomes a critical task.

Likewise, altmetrics present a pronounced disciplinary bias. Topics with high media impact and immediate applicability, such as a medical breakthrough or a natural disaster, naturally dominate the digital space. Conversely, disciplines that operate on longer, more specialized cycles of debate, such as theoretical philosophy or pure mathematics, naturally generate a much fainter altmetric footprint. This does not mean their social impact is less; rather, it manifests in more subtle, less quantifiable ways through these channels, which may penalize them in evaluations that prioritize these indicators.

Finally, the lack of methodological standardization impedes comparability. Platforms such as Altmetric.com and PlumX use their own algorithms, weight sources differently (e.g., views on Wikipedia versus mentions in the press), and define their “score” in unique ways. The absence of a unified protocol means that the same article can have radically different altmetric scores across platforms, making it difficult to create reliable benchmarks and undermining consistency in evaluation.

Ultimately, altmetrics should not replace traditional indicators, but rather complement them. Their maximum value is achieved when they are approached with a critical eye, with an understanding of their biases, and used to tell a richer, more nuanced story about the dissemination of knowledge in society, rather than as a simple, definitive number.

Integrated case example: a study on vaccines

Bibliometrics:

- Analysis: 5000 articles on vaccines in Scopus.
- Finding: 60 % of publications come from the US, China, and the UK.

Scientometrics:

- Analysis: patterns of collaboration between countries.
- Finding: international teams produce 40 % more patents.

Altmetrics:

- Analysis: mentions on Facebook and in the news.
- Finding: articles with summaries in plain language have three times more public impact.

Three tools, one purpose

These disciplines are not mutually exclusive, but complementary:

- Bibliometrics answers “Who cites whom?”
- Scientometrics explains “Why does science grow this way?”
- Altmetrics reveals “How does society use this knowledge?”

“Using only bibliometrics is like measuring an ocean with a glass: you need altmetrics to see the waves and scientometrics to understand the currents.”

1.4 Practical applications: scientific evaluation and research policies

From data to decisions

Bibliometrics transcends mere citation counting to become a **strategic tool** in knowledge management. This section explores how governments, universities, and funding agencies use bibliometric indicators to **evaluate science** and **design policies**, with concrete examples and ethical controversies.

Scientific evaluation

Chapter 5 will discuss evaluation indices and methods in greater depth. Here are some of their uses.

Evaluation of researchers

The evaluation of researchers using indicators presents a complex balance between the objective measurement of scientific impact and the preservation of academic integrity. Among the most widely used indicators are the h-index and its variants (g and m), which aim to capture a researcher’s productivity and impact simultaneously, though they have inherent limitations. Field-normalized citation emerges as a more refined tool, allowing fair comparisons between disciplines with different publication and citation dynamics. Complementarily, the percentage of publications in the top-citation percentile (top 10 %) provides a valuable perspective for identifying truly transformative research in each area of knowledge.

However, these evaluation systems are not without significant controversy. Even more problematic is the fetishism of the h-index, a phenomenon that has had unintended consequences in the scientific ecosystem: on the one hand, it discourages risky and disruptive research, which tends to have longer maturation cycles; on the other, it promotes a culture of incremental publication to the detriment of fundamental but less frequent contributions.(12) These tensions between quantitative metrics and actual scientific quality continue to generate intense debates about how to evaluate research merit without sacrificing epistemological diversity and radical innovation.

Real-life example:

> *Leiden University (Netherlands) uses normalized citations to avoid bias against researchers in fields with low citation rates (e.g., mathematics vs. biomedicine).*[13]

Journal evaluation

Evaluating the prestige and influence of academic journals is a cornerstone of bibliometrics, but it has evolved to overcome the exclusive reliance on univocal metrics. The Journal Impact Factor (JIF), calculated by Clarivate Analytics, has been the dominant indicator for decades. It measures the frequency with which articles from a journal published in two years are cited in a given year. However, its methodology has been the subject of substantial criticism: it has an inherent bias toward English-language and natural science journals, marginalizing high-quality publications in the humanities, social sciences, or regional journals, whose citation dynamics are slower.

In response to these limitations, alternative indicators have emerged that offer more nuanced perspectives. The SCImago Journal Rank (SJR), based on the Scopus database, introduces a key concept: prestige transfer. Not all citations are equal; a citation from a highly prestigious journal carries more weight than one from a lesser publication. This approach, inspired by Google's algorithm, helps identify journals that are influential leaders within their specific niches, even if their total number of citations is not the highest.

Complementarily, CiteScore (also from Elsevier/Scopus) expands the time window and the type of documents considered in the calculation. It includes not only research articles, but also reviews, conference proceedings, book chapters, and notes, making it particularly relevant for disciplines such as engineering or computer science, where conference proceedings are of seminal value. This greater inclusiveness provides a more representative view of a journal's impact on its entire disciplinary ecosystem.

Institutional evaluation

The bibliometric evaluation of universities and research centers has transcended simple publication counts to adopt advanced metrics that reflect quality, internationalization, and scientific leadership.

Two indicators are particularly revealing:

- Degree of International Collaboration: calculated as the percentage of an institution's publications that have foreign co-authors. A high index is a strong indicator of the institution's integration into global knowledge networks, its ability to attract international talent, and its access to large-scale projects.
- Research Excellence Index: measures the percentage of an institution's articles that are among the top 10 % most cited in the world in their respective fields. This metric is not limited to quantifying production, but identifies an institution's ability to generate cutting-edge, high-impact research, that is, to be at the forefront of knowledge.

Together, these instruments, for both journals and institutions, facilitate a crucial transition: moving from a culture of quantity to a culture of impact and excellence, provided they are interpreted with an awareness of their contexts and limitations.

1.4.1. Research policies

As demonstrated in previous sections, bibliometrics transcends its descriptive function to become a fundamental tool in the governance of science. By providing solid, quantitative evidence, it enables governments and institutions to design more innovative, more strategic, and more equitable research policies. Far from being a mere post-hoc evaluation tool, bibliometrics informs a priori decision-making, guiding the distribution of resources, fostering high-impact collaborations, and accelerating the transition to more open and collaborative science. Below are three key applications in scientific policy management.

Funding distribution: investing strategically, not by inertia

Funding agencies have left behind the era of allocations based solely on intuition or historical prestige. Today, bibliometric analyses offer a dynamic map for investing in what really matters. Through techniques such as science mapping, it is possible to identify emerging areas with exponential growth in publications and citations, pinpointing where an investment will yield the maximum return. Complementarily, citation network analysis and the identification of knowledge gaps reveal neglected scientific topics that are nevertheless crucial for interdisciplinary advancement or for addressing urgent social challenges, allowing funds to be proactively directed toward these gaps.

Designing collaboration programs: connecting global talent

Contemporary science is inherently collaborative. Bibliometrics acts as a catalyst for strategically designing these alliances. Co-authorship analysis identifies institutions and countries with complementary research interests and synergistic strengths, paving the way for joint project calls. The study of scientific mobility, tracking the rate of researchers migrating between regions, provides crucial data for talent retention and attraction policies. The fact that approximately 35 % of African scientists specializing in AI work in the US is a well-established pillar of science policy studies. Numerous organizations, including the OECD and UNESCO, cite it.(14)

Open science policies: measuring to democratize

As discussed previously, open access is a transformative force. Bibliometrics provides the indicators needed to measure progress toward this goal and design effective policies that democratize knowledge. Platforms such as the EU's Open Access Panel, mentioned above, allow real-time monitoring of the percentage of open-access publications within a country or institution, the predominance of different routes (gold, green), and the gaps across disciplines and regions. This data is essential for adjusting mandates, negotiating with publishers, and ensuring that the transition to open science is inclusive and reduces, rather than widens, inequalities in access to knowledge.

Bibliometrics is no longer a simple mirror reflecting scientific activity. By providing evidence for strategic decision-making, it has become a driving force that not only describes science but also transforms it, helping to build a more efficient, connected, and open research ecosystem.

1.4.2. Bibliometrics as a social barometer

Bibliometrics is far from being a neutral tool: it reflects the values, priorities, and biases inherent in the contemporary scientific ecosystem. Its true transformative power lies in three critical dimensions:

First, it serves as **a compass for optimizing scarce resources**. Quantifying the demonstrable impact of research, whether through citations, technology transfer indicators, or public policy metrics, allows funding to be allocated based on evidence rather than intuition or academic tradition. This approach is particularly valuable in contexts where every R&D investment must be justified to societies that demand tangible returns.

At the same time, well-applied bibliometrics can **democratize evaluation processes**. Compared to traditional systems based on networks of influence or institutional reputation, standardized metrics offer, in theory, a common and reproducible language.

However, its most revolutionary potential lies in its ability to **expose and combat structural inequalities**. Bibliometric analyses have starkly exposed the geographical concentration of scientific production, persistent gender gaps (female researchers systematically receive fewer citations in fields such as AI or theoretical physics), and the metric disregard for disciplines such as the humanities or social sciences.

The ethical challenge is to use these diagnoses not to perpetuate existing hierarchies, but to design corrective policies, such as differentiated evaluation criteria by discipline or funding quotas for underrepresented regions.

This tension between reproducing and transforming the status quo makes bibliometrics a crucial battleground for defining what kind of science we want to build: one that rewards only immediate productivity, or one that also values epistemological diversity and long-term social impact. This guide will provide the tools to navigate this dilemma with methodological rigor and critical awareness.

“Metrics are like a scalpel: in expert hands they save (scientific) lives; in inexperienced hands, they mutilate careers.”

1.5. Ethics and good practices in bibliometrics

1.5.1. Ethical risks: when numbers reinforce inequalities

Bibliometrics, with its power to quantify, has brought an appearance of objectivity to scientific evaluation. However, it is essential to recognize that these indicators do not operate in a neutral vacuum. On the contrary, they are imbued with the biases and power dynamics that pre-exist in the academic system, and can perpetuate and amplify them. This section outlines the main ethical risks of an uncritical application of bibliometrics and explores emerging strategies to build a fairer, more representative scientific evaluation.

The “Matthew Effect” in Science

This phenomenon, named by sociologist Robert K. Merton after a biblical passage, starkly describes how an initial advantage attracts cumulative advantages. In science, this manifests in Price’s Law, which posits that a minority (about 20 % of researchers) receives the vast majority (approximately 80 %) of citations and recognition.[15]

This virtuous cycle for a few becomes a trap for many. Brilliant young researchers, institutions in developing countries, and scientists in marginal fields struggle hard to achieve visibility, even when they produce work of the highest quality. By rewarding mainly those already well known, the system can stifle innovation emerging from the margins.

The rigid application of these indicators can also distort incentives, rewarding rapid productivity and predictable citability over risky, far-reaching research or research with social impact that is not immediately quantifiable. There is also a risk of excluding valuable forms of knowledge, such as local knowledge or contributions in languages other than English, which do not fit into the formal channels of indexed publication.

Against this backdrop, the movement for responsible scientometrics is gaining momentum, advocating for the contextualized use of indicators, qualitative peer review, and the development of metrics that capture the diversity of contributions to the knowledge ecosystem. Bibliometrics should be a servant of science, not its master.

Bibliometrics also reveals and, at times, reinforces deep structural inequalities. A clear example is gender bias, as mentioned above.

The crisis of quantitative evaluation

The rigid application of bibliometric indicators has led to profound distortions in scientific evaluation, in which arbitrary figures can outweigh actual intellectual merit. This crisis manifests when promotion committees establish inflexible numerical barriers, such as requiring an Impact Factor greater than 20, which distort the very meaning of evaluation by ignoring the intrinsic quality, originality, and concrete contribution of a piece of research. These abuses in the use of metrics, where a number obscures the real value of the work, motivated the creation of global initiatives such as the DORA (San Francisco Declaration on Research Assessment), which seek to restore expert and qualitative judgment as the central axis of scientific evaluation.(16)

The two pillars of responsible bibliometrics

DORA Declaration

"Evaluate science for its content, not its container."

The DORA Declaration (San Francisco Declaration on Research Assessment), established in 2013, marks a turning point in scientific evaluation by proposing a paradigm shift: evaluating research on its actual content rather than the prestige of the journal in which it is published. This initiative emerged as a critical response to the overvaluation of journal Impact Factors (IFs), promoting a more holistic and fair assessment of research instead. Its fundamental principles include the express prohibition of using IF as a measure of quality for individual researchers, the recognition of the diversity of scientific products, from data sets and software to outreach activities, and the requirement for maximum transparency in the calculation methods and data sources used for any evaluation.(16)

Its widespread adoption reflects a growing consensus on the need to reform scientific evaluation systems, shifting the focus from superficial metrics to a qualitative, context-based assessment of contributions to knowledge. The declaration has also promoted the development of new evaluation practices that recognize the social value of research and encourage open science, paving the way for a more equitable academic ecosystem focused on the actual impact of scientific work.

Leiden Manifesto

"10 Commandments for the Use of Indicators"

The Leiden Manifesto (2015), which proposes a set of fundamental principles for the appropriate use of bibliometric indicators, emphasizes the importance of adjusting analyses according to the discipline in which they are applied.(17)

According to this principle, an h-index of 12 could be considered excellent in the field of philosophy but low in biology, due to inherent differences in publication and citation dynamics across different areas of knowledge. Furthermore, it stresses the need to prioritize quality over quantity when evaluating the impact of research. Thus, a genuinely innovative and revolutionary article should carry more weight than several trivial studies that only contribute to the accumulation of citations without generating real progress in the discipline. Finally, the manifesto emphasizes the importance of auditing the databases used for these analyses, as many, such as Scopus, may overlook a large part of academic production, especially in the humanities, where 80 % of relevant books may be excluded from analysis.

Together, these principles aim to make a more rigorous and thoughtful use of bibliometric indicators to avoid distortions in the evaluation of science.

The seven deadly sins of bibliometrics

Bibliometrics was created to illuminate the paths of science, but its excessive application has created new labyrinths. What began as a guidance tool has, in many cases, become an end in itself, distorting the fundamental values of research. Identifying these “deadly sins” of bibliometric evaluation is the first step toward recovering its original meaning: to serve science, not dominate it.

The ***fetishization of the impact factor*** is the most obvious deviation. We have turned a number created to aid in journal selection into an unquestionable oracle. This idolatry has had paradoxical consequences: the more obsessed we are with the impact factor, the less we pay attention to what really matters, the intrinsic quality of the research. The solution is not to abandon metrics, but to choose them wisely, opting for more nuanced indicators such as SJR or CiteScore, which offer a more contextual view and are less susceptible to manipulation.

Inflated self-citations tell us a sad truth: in their eagerness to climb the ranks, some researchers have turned citations into a bargaining chip rather than genuine intellectual recognition. When references cease to be an academic dialogue and become a positioning strategy, the very essence of science as a collaborative endeavor is undermined. Setting reasonable limits, such as the 20 % proposed by the prestigious Leiden Center, is not just a technical issue: it is a reminder that integrity must prevail over opportunism.

Ignorance of disciplinary contexts is equivalent to judging a fish by its ability to climb trees. The humanities, social sciences, and arts have rhythms, formats, and traditions of communication that do not fit into the molds designed for the natural sciences. Demanding that they play by the same rules is not only unfair but also scientifically incorrect. True excellence is measured within each disciplinary ecosystem, not through forced comparisons that only generate frustration and homogenization.

The ***thoughtless use of institutional rankings*** has created a “wealth begets wealth” effect that deepens global gaps. Institutions in the Global South, no matter how talented they may be, struggle against an evaluation system that rewards precisely the resources they lack. Rankings do not only measure quality, but they also measure historical privileges and structural advantages. That is why we need indicators that know how to contextualize, that understand that excellence in Nairobi is not, nor should it be, the same as excellence at Harvard.

Excluding books as a source of evaluation is equivalent to amputating a large part of the most profound thinking in the humanities and social sciences. While we reward short articles and immediate results, we penalize thoughtful reflection, meticulous study, and work that takes years to mature. Incorporating tools such as Google Scholar that recognize these formats is not a concession; it is an act of epistemological justice.

The ***metric of obesity*** has led us to believe that more data means better evaluation. But experience shows us the opposite: when everything is measured, nothing is understood. The endless proliferation of indicators has not brought us closer to the truth; it has distracted us from it. Elegance lies in simplicity: five well-chosen indicators can tell us much more than fifty misinterpreted ones.

Algorithmic opacity is the most insidious challenge. We evaluate using tools whose criteria we do not know, relying on black boxes that may hide the very biases we claim to combat. Transparency here is not just a virtue; it is a condition for survival. Without public audits, without open scrutiny, bibliometric evaluation becomes an unquestionable dogma.

Correcting these excesses is not a step backward, but a necessary maturation. It invites us to build a bibliometrics on a human scale, rigorous but comprehensive, ambitious but humble. One that knows that behind every number there is a researcher, behind every ranking there is an institution, and behind every metric there is an ethical decision about what science we want and for whom.

Framework for Good Practice (UNESCO, 2021)

In response to growing distortions in scientific evaluation systems, UNESCO has established a framework of good practices to rebalance how we understand and value knowledge production. This framework, the result of global consensus, proposes concrete transformations in the three pillars of the academic ecosystem: evaluation, authorship, and publishing.[18]

For evaluators, the framework proposes a balanced formula that combines traditional metrics with qualitative peer review at a 70/30 ratio. This approach recognizes the value of bibliometric data but places expert judgment at its center. The novelty lies in how metrics are integrated: it urges the systematic incorporation of altmetrics to capture the social impact of research, thus recognizing that the value of knowledge transcends academic citations and manifests itself in its ability to influence public policy, transform social practices, or enrich cultural debate.

In the area of authorship, the framework addresses two particularly harmful practices. On the one hand, it discourages “salami slicing,” the tendency to artificially segment research into as few publications as possible, reminding us that intellectual integrity must take precedence over numerical productivity. On the other hand, it establishes an ethical threshold for self-citation, recommending that it should not exceed 15 % of total citations. This limit does not seek to restrict the coherent construction of a research career, but rather to ensure that work is validated through recognition by the academic community as a whole, thus preserving objectivity in evaluation.

Finally, the framework turns its attention to publishers, key players in preserving scientific integrity. It urges them to practice radical transparency in their editorial selection algorithms, making public the criteria that determine what is published and what is not. At the same time, it requires them to eradicate “guest authorship,” the inclusion of ghost authors for reasons of convenience or prestige, through the rigorous implementation of authorship criteria based on actual contributions. These measures seek to restore confidence in a publishing system whose credibility has been threatened by opaque practices.

Taken together, these guidelines represent a shift toward a more humane and transparent science. They do not merely correct abuses, but propose a new philosophy of evaluation where quality, integrity, and social impact are intertwined to create a more robust scientific ecosystem that is ultimately more faithful to its mission of serving society.

Case Study: How to Implement DORA in a Department

Step 1: Training in responsible metrics (4 hours).

Step 2: Replace FI with:

- Normalized h-index by discipline.
- Percentage of articles in the top 10 % of citations.

Step 3: Introduce qualitative assessment through:

- 2 external peer reports.
- 1 social impact letter.

A real case:

Result at UC Davis (2023):

- 32 % more women promoted.
- 25 % increase in interdisciplinary research.

Best practices in 4 steps

This guide recommends:

1. Audit biases: use *the Gender Citation Gap Analyzer* (Python tool) to detect disparities.
2. Triangulate metrics: E.g.: if a researcher has few citations but a high impact on policy (measured via *Overton*), evaluate both.
3. Train evaluators: Workshops to avoid judgments based on prejudice (e.g., dismissing journals in Spanish).
4. Demand transparency: Universities should publish their bibliometric criteria.

The goal is not to abandon data, but to use it to amplify silenced voices and reward science with social significance.

Towards a culture of fair evaluation

Ethical bibliometrics does not reject numbers; instead, it rejects their simplistic use. As *Ludo Waltman* (co-author of the Leiden Manifesto) summarizes:

"When metrics become ends, they corrupt science. When they are means, they improve it."

Chapter 2 will explore the history of bibliometrics, from Garfield's citation indexes to the use of artificial intelligence.

Recap

- Bibliometrics is the discipline that applies quantitative methods to measure the production, dissemination, and impact of scientific knowledge.
- It emerged as part of information and documentation sciences, linked to the evaluation of research activity.
- Its purpose is to evaluate and understand the dynamics of science, not just to quantify it.
- It is based on the analysis of publications and citations, which are considered measurable traces of the advancement of knowledge.
- Modern bibliometrics combines statistics, computer science, and the theory of scientific communication.
- It provides objective, comparable indicators that are fundamental for scientific policies and university management.

- The main objects of analysis are authors, institutions, journals, countries, and topics.
- Fundamental indicators include: productivity (number of publications), impact (number of citations), and collaboration (co-authorship).
- There are composite indicators such as the h-index, g-index, impact factor, and SJR.
- Bibliometrics does not replace qualitative evaluation; rather, it complements it with verifiable empirical evidence.
- Its historical development began in the 1960s with Garfield and the Science Citation Index.
- It has evolved into scientometrics and altmetrics, expanding its scope to include network analysis and digital media.
- It is an essential tool for measuring visibility, collaboration, and scientific influence.
- It is applied in institutional ranking, journal evaluation, trend analysis, and technology watch.
- Bibliometrics is helpful for both evaluators (agencies, committees, universities) and individual researchers.
- Its results should be interpreted with caution: the figures do not always reflect scientific quality or relevance.
- Correct interpretation requires disciplinary, temporal, and linguistic contextualization.
- Its limitations include coverage, language, and self-citation biases.
- Metrics should be used ethically and transparently, avoiding “number fetishism.”
- When applied correctly, bibliometrics strengthens accountability, scientific planning, and open science.

Self-assessment questions

1. How is bibliometrics defined, and what is its primary purpose?
2. What elements constitute the empirical basis of bibliometric analysis?
3. What disciplines converge in the modern development of bibliometrics?
4. What are the three main types of bibliometric indicators?
5. What is the difference between simple indicators (such as the number of citations) and composite indicators (such as the h-index)?
6. Why should qualitative evaluation complement bibliometrics?
7. What role did Eugene Garfield play in the history of bibliometrics?
8. What biases or limitations can affect bibliometric results?
9. In which institutional settings is bibliometrics currently applied?
10. What ethical principles should govern the use of bibliometric indicators?

BIBLIOGRAPHY

1. Moed HF. Citation analysis in research evaluation. Dordrecht: Springer; 2005. doi:10.1007/1-4020-3714-7

2. Bornmann L, Daniel HD. What do citation counts measure? A review of studies on citing behavior. J Doc. 2008;64(1):45–80. doi:10.1108/00220410810844150

3. Thelwall M. Web indicators for research evaluation: A practical guide. San Rafael (CA): Morgan & Claypool Publishers; 2016. doi:10.2200/S00733ED1V01Y201602ICR048

4. Leydesdorff L, Milojević S. Scientometrics. In: Wright JD, editor. International encyclopedia

of the social & behavioral sciences. 2nd ed. Amsterdam: Elsevier; 2015. p. 322–7. doi:10.1016/B978-0-08-097086-8.85017-7

5. Sugimoto CR, Larivière V. Measuring research: What everyone needs to know. Oxford: Oxford University Press; 2018. https://global.oup.com/academic/product/measuring-research-9780190640125

BIBLIOGRAPHIC REFERENCES

1. Lim WM, Kumar S. Guidelines for interpreting the results of bibliometric analysis: a sensemaking approach. Glob Bus Organ Excell. 2024;43(2):17–26.

2. Yepes-Nuñez JJ, Urrútia G, Romero-García M, Alonso-Fernández S. Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. Rev Esp Cardiol. 2021;74(9):790–9.

3. Passas I. Bibliometric analysis: the main steps. Encyclopedia. 2024;4(2):1014–25.

4. Maddi A, Da Silva JAT. Beyond authorship: analyzing contributions in PLOS ONE and the challenges of appropriate attribution. J Data Inf Sci. 2024;9(3):88–115.

5. Kojaku S, Livan G, Masuda N. Detecting anomalous citation groups in journal networks. Sci Rep. 2021;11(1):1–11.

6. Diprose JP, Hosking R, Rigoni R, Roelofs A, Chien TY, Napier K, et al. A user-friendly dashboard for tracking global open access performance. J Electron Publ. 2023;26(1):17–61.

7. Oza A. Citations show gender bias — and the reasons are surprising. Nature. 2023 Dec 22.

8. Zhang X, Li J, Gu Z, Gao X. How does language learning contribute to individual growth in a multilingual world? A systematic review. J Multiling Multicult Dev. 2025 Sep 30.

9. Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 2016;106(1):213–28.

10. García-Villar C, García-Santos JM. Indicadores bibliométricos para evaluar la actividad científica. Radiología. 2021;63(3):228–35.

11. Peralta González MJ, Maylín I, Guzmán F, Gregorio O, Ii C. Criterios, clasificaciones y tendencias de los indicadores bibliométricos en la evaluación de la ciencia. Rev Cub Inf Cienc Salud. 2015;26(3):290–309.

12. Rousseau R, Hu X. Metrics: a fetish for high-profile journals. Nature. 2012;490(7420):343.

13. Waltman L, Calero-Medina C, Kosten J, Noyons ECM, Tijssen RJW, Van Eck NJ, et al. The Leiden ranking 2011/2012: data collection, indicators, and interpretation. J Am Soc Inf Sci Technol. 2012;63(12):2419–32.

14. Directorate for Education and Skills. Paris: Organisation for Economic Co-operation and Development (OECD); c2025. https://www.oecd.org/en/about/directorates/directorate-for-education-and-skills.html

15. De Solla Price DJ. Networks of scientific papers. Science. 1965;149(3683):510–5.

16. The American Society for Cell Biology. San Francisco Declaration on Research Assessment (DORA). 2012 Dec 16. https://sfdora.org/

17. Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I. Bibliometrics: the Leiden manifesto for research metrics. Nature. 2015;520(7548):429–31.

18. UNESCO. Recomendación sobre la ciencia abierta. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura; 2021.

Chapter 2 / Capítulo 2

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch02

Brief Historical Overview / Breve Recuento Histórico

In 1955, a young chemist named Eugene Garfield published an article entitled *"Citation Indexes for Science,"* in which he proposed a system for tracking the impact of research. No one imagined that this idea would revolutionize science, giving rise to modern bibliometrics.[1]

2.1. The Analog Era (1960-1990): The Foundations

Science Citation Index (SCI) - 1960

In 1960, the Institute for Scientific Information (ISI), founded by Eugene Garfield, revolutionized scientific evaluation with the launch of the Science Citation Index (SCI), the first large-scale commercial citation index. Originally published in print, this system made it possible, for the first time, to systematically track how scientific articles were linked through their bibliographic references.[2]

The SCI was based on a simple but powerful principle: *"citations are connections of knowledge."* Its methodology included tracking cross-references among 600 selected scientific journals (mainly from the US and Europe), reverse indexing (instead of just listing authors or topics, the SCI allowed users to search for articles that cited a particular work, revealing its subsequent influence), and multidisciplinary coverage (although focused on the natural sciences, it laid the foundation for future indexes in the social sciences and humanities).

The SCI transformed the way scientific impact was measured. Before the SCI, productivity was measured by the number of publications. The index introduced the idea that an article could be influential even if its author published little (e.g., Watson and Crick's paper on DNA had few previous publications but changed biology). It identified highly cited works that defined entire fields (e.g., Miller's 1953 article on the origin of life). Despite its limitations, such as Anglophone bias, Western focus, and restricted access to privileged institutions, its influence endures through contemporary platforms such as Web of Science. At the same time, Google Scholar and Scopus have adopted their fundamental cross-citation logic.[3]

The SCI didn't just measure science: it made it more transparent and connected. Its history reminds us that even the most disruptive tools must be used with critical awareness.

Fun fact:

> *The original SCI occupied five linear meters of shelving, and only elite institutions such as Harvard could afford it.*[4]

Fundamental Laws (1970s-1980s). The mathematical pillars of bibliometrics

During the 1970s and 1980s, bibliometrics consolidated its scientific rigor through the validation and widespread application of two empirical laws formulated decades earlier: Lotka's Law (1926) and Bradford's Law (1934). These laws, although initially developed to describe patterns in scientific literature, became essential tools for understanding researcher productivity and the distribution of knowledge in academic journals.

Lotka's Law: inequality in scientific productivity

Lotka's Law, formulated by Alfred J. Lotka in 1926 but popularized during the scientometric boom of the 1970s, is one of the fundamental findings on the unequal distribution of scientific productivity. This principle states that approximately 10 % of researchers produce 50 % of academic publications, revealing a pattern of concentration of scientific output that transcends disciplines and historical periods. The law represents the specific application of the Pareto principle to the scientific domain, confirming that knowledge production follows an asymmetric distribution in which a few actors generate most of the research results.[5]

Contemporary empirical evidence consistently validates this unequal distribution across multiple fields of knowledge. This distribution enhances the "Matthew effect" in science, where established researchers with access to collaborative networks, institutional resources, and accumulated symbolic capital tend to publish more frequently, thus reinforcing their initial competitive advantages.

The practical applications of Lotka's Law in academic evaluation are numerous and significant. It enables the identification of highly productive researchers for hiring, promotion, or funding-allocation decisions, providing a quantitative reference point for evaluating exceptional performance. Simultaneously, it serves as a diagnostic tool for identifying structural inequalities within specific scientific systems, informing policies to democratize opportunities for publication and academic visibility.

However, the law has significant methodological limitations in contemporary contexts. Its applicability is more robust in STEM disciplines than in the humanities, where single authorship and diverse publication formats that escape traditional metrics predominate. Nor did it anticipate the explosion of massive co-authorship characteristic of large-team science, where publications in genomics or experimental physics can include thousands of authors, redistributing the dynamics of individual productivity. These limitations underscore the need to contextualize the law within the paradigmatic changes in scientific authorship and collaboration practices of the 21st century.

Bradford's Law: the core of key journals

Bradford's Law (1943) is a fundamental principle of bibliometrics that describes the uneven distribution of relevant scientific literature across academic journals. It establishes that a small core of periodicals concentrates the majority of significant articles in any field of knowledge, while a more extensive periphery hosts scattered contributions. This pattern of concentration reflects the specialized structure of communication in contemporary science, where specific journals serve as privileged channels for disseminating high-impact research.[6] Empirical evidence consistently confirms this phenomenon of concentration across multiple disciplines.

The practical applications of this law are particularly valuable for information resource management. Academic libraries use Bradford zone analysis to optimize subscriptions, focusing limited resources on the core of journals that maximize access to relevant literature. Simultaneously, the identification of "desert zones," fields where knowledge is widely dispersed among numerous sources, alerts us to the need for more comprehensive search strategies, especially in interdisciplinary areas or specific cultural studies.

However, Bradford's Law faces significant limitations in the contemporary scientific ecosystem. It has a marked Anglophone bias, with most articles published in large databases being in English, marginalizing valuable contributions in other languages. In addition, the emergence of open science and preprint repositories is substantially transforming these patterns of concentration. In artificial intelligence, for example, arXiv has disrupted traditional publishing by creating an alternative channel of dissemination that competes with established journals, demonstrating the growing fluidity of scientific communication patterns.

This evolution toward more distributed models suggests that, although the principle of concentration remains relevant, its concrete manifestations are rapidly changing. Contemporary bibliometrics must develop tools capable of capturing these new dynamics while maintaining the analytical utility of Bradford's original concept for understanding the structure of scientific communication.

General limitations of these laws

Although revolutionary, both laws have problems in the digital age:

1. Biased coverage: they were based on data from the US and Europe, ignoring scientific output from Asia, Africa, and Latin America. Example: Bradford's Law does not predict well the core of journals in African agronomy, where research is published in local journals.
2. Manual processing: in the 1970s and 1980s, counts were done by hand, which led to errors (e.g., duplication of authors with similar names).
3. Current context: today, algorithms such as those used by Scopus or Dimensions allow for more dynamic analysis, but the laws remain the theoretical basis.

Why do they still matter in the 21st century?

These historical bibliometric laws remain surprisingly relevant in the 21st-century digital scientific ecosystem, demonstrating that the fundamental patterns of academic communication transcend technological change. Lotka's Law continues to inform the development of modern evaluation systems, where the h-index and its variants incorporate its understanding of the uneven distribution of scientific productivity. This perspective contextualizes individual metrics within broader systemic patterns, preventing simplistic interpretations of research performance and recognizing the inherently asymmetrical nature of knowledge production.

In the realm of scientific policy, Bradford's Law provides a crucial analytical framework for navigating the exponential expansion of academic communication. Its principle of concentration helps identify predatory journals operating outside the legitimate cores of each discipline, offering an objective criterion for distinguishing reliable communication channels. This application is particularly valuable in open access contexts, where the proliferation of questionable publishers requires robust mechanisms to ensure the integrity of scientific communication.

The most profound influence of these laws is manifested in the algorithmic foundations that underpin contemporary digital tools. The PageRank algorithm of Google Scholar, for example, computationally implements Bradford's principle by assigning greater weight to citations from sources considered "nuclei" of academic authority. Simultaneously, scientific literature recommendation and discovery systems incorporate insights derived from Lotka to prioritize content from highly productive and influential researchers.

The continued validity of these principles demonstrates that, although scientific communication technologies have undergone radical transformations, the underlying patterns of knowledge production and dissemination maintain observable structural regularities. This continuity makes classical bibliometric laws essential tools for understanding both the continuities and transformations in contemporary scientific dynamics, providing an interpretive framework for navigating the complexity of today's research ecosystem.

2.2. The Digital Revolution (1990-2010): expansion and criticism

The Era in Which Bibliometrics Became Global (and Controversial)

The 1990s marked the beginning of **digital bibliometrics**, radically transforming how science is measured and managed. With the advent of the Internet, citation indexes moved away from print formats and adopted online platforms, expanding their reach but also generating new ethical and methodological challenges.

Web of Science (WoS) – 1997: The SCI goes digital

It was the digital version of *the Science Citation Index (SCI)*, launched by the **Institute for Scientific Information (ISI)**. For the first time, it enabled **real-time searches** and advanced citation analysis. It grew from 600 journals (in print) to 8,000 indexed journals, including social sciences (Social Sciences Citation Index) and arts (Arts & Humanities Citation Index). It introduced features such as *Citation Reports* (to calculate the impact of authors) and *the Journal Impact Factor (JIF) online* (previously available only in the printed *Journal Citation Reports*).(7)

The JIF, created in 1975 to evaluate journals, began to be misapplied to individual researchers. Universities require publication in “Q1 journals” for hiring, ignoring the actual quality of the articles.

Scopus (2004): the competitor that challenged the hegemony

Developed by Elsevier as a direct alternative to Web of Science (WoS), Scopus broadened the horizon of scientific indexing by incorporating a greater number of non-English-language journals, with a strong focus on European and Asian publications. Among its most significant contributions, it introduced author profiles and pioneered systems for name disambiguation that anticipated tools such as ORCID. It also incorporated alternative metrics such as CiteScore, conceived as a complementary indicator to the Journal Impact Factor (JIF).

However, Scopus was not without its critics. It has been accused of commercial bias, favoring Elsevier-associated publications in its rankings. Furthermore, its refusal to index preprints put it at a disadvantage compared to more open systems such as Google Scholar, as it excluded a significant portion of informal or “gray” scientific literature from its database.

Google Scholar (2004): democratic disruption

The launch of Google Scholar represented a profound transformation in access to scientific knowledge. Its automatic indexing system allowed it to incorporate not only peer-reviewed articles, but also preprints, such as those from arXiv, books, theses, and technical documents, many of which had remained off the radar of traditional databases. Unlike WoS and Scopus, Google Scholar was free, democratizing access to global scientific literature. Its ability to capture content in traditionally marginalized languages and formats, such as studies in Spanish from Latin America, gave it unprecedented coverage. In addition, with the development of Google Scholar Citations, researchers could create public profiles to showcase their academic output and personal metrics.

However, this openness also generated controversy. One of the main questions has been the opacity of its algorithms, as there is no information on how results are ranked, leading to strategies to manipulate the visibility of specific works through the intentional use of keywords. On the other hand, the quality of the indexed documents is a matter of debate, as Google Scholar includes predatory journals and non-peer-reviewed materials.

Impact and criticism of this revolution

The emergence of platforms such as Scopus and Google Scholar marked a turning point in the way scientific knowledge is accessed, evaluated, and disseminated. Among their main legacies is the globalization of science, by giving visibility to research from historically marginalized regions, such as Latin America, Asia, and Africa, beyond the traditional centers of academic production in the United States and Europe. In addition, the massive digitization of citations has enabled the development of new metrics, with the h-index, proposed by Hirsch in 2005, among the most influential for simultaneously measuring a researcher's productivity and impact.

However, structural problems persist that these platforms have not been able to resolve. Despite accumulated criticism, the impact factor (JIF) continues to be used as a decisive criterion in evaluation and funding processes, as is the case with national agencies such as ANID in Chile.[8] This persistence reinforces a quantitative logic that often ignores the quality or context of the knowledge produced. On the other hand, the digital divide remains a significant obstacle: many institutions in low- and middle-income countries cannot afford access to commercial databases such as WoS or Scopus, forcing them to rely on free alternatives such as Google Scholar, with all the limitations and risks that this implies in terms of reliability. Finally, the rise of digital tools has given rise to new forms of metric manipulation, such as citation stacking, a practice that consists of creating circles of fraudulent citations to artificially inflate the impact of specific publications, thus distorting academic evaluation processes.

2.3. The Modern Era (2010-Present): Complexity and Democratization

Altmetrics (2010)

Starting in 2010, bibliometrics began to expand into broader forms of scientific impact assessment, giving rise to altmetrics. This approach seeks to measure impact beyond the counting of academic citations, incorporating indicators such as mentions in social media, the media, blogs, Wikipedia, and public policy documents. The launch of Altmetric.com in 2011 marked a turning point, offering tools to track the digital circulation of scientific articles in real time.

Open Source Bibliometrics (2017)

In 2017, the democratization of bibliometrics took a significant leap forward with the emergence of open-source tools, which facilitated access to advanced analysis without the high costs of commercial platforms. Among the most notable are Bibliometrix, a R package for sophisticated bibliometric analysis, and VOSviewer, a widely used tool for visualizing co-authorship, co-citation, and keyword networks. These platforms enabled researchers at institutions with limited resources to access robust analytical methodologies, contributing to the decentralization of bibliometric knowledge.

Artificial Intelligence and Text Mining (2020)

In the 2020s, the integration of artificial intelligence revolutionized the way scientific literature is processed and analyzed. Models such as GPT-4 enable automated, efficient summarization of trends from millions of abstracts, allowing synthetic analysis of entire fields of research in record time. At the same time, platforms such as Dimensions.ai began to integrate databases that connect scientific publications with patents, grants, and research results, offering a much more comprehensive perspective on the cycle of knowledge production and transfer. This convergence between AI, text mining, and open science has expanded the scope and depth of contemporary bibliometrics, positioning it as a key tool in scientific and public policy decision-making.

Current challenge:

> *How to prevent AI from generating "zombie articles" (well-cited texts with no real substance).*

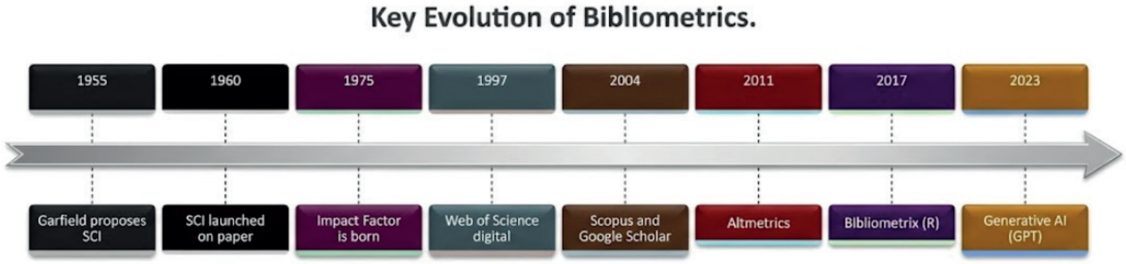


Figure 2.1. Visual timeline

2.4. Historical lessons

1. From analog to digital: Technological leaps have democratized access but introduced new biases (e.g., information overload).
2. From quantitative to qualitative: Criticism of IF led to responsible metrics (DORA, Leiden).
3. From academic to social: Altmetrics broadened the notion of scientific impact.

“Bibliometrics is no longer just about counting citations, but understanding how knowledge flows—and sometimes stagnates—in society.”

Section II will explore the most commonly used tools for constructing bibliometric analyses and how to obtain data for bibliometric studies practically.

Recap

- Bibliometrics has its origins in the first half of the 20th century, when the first attempts to quantify scientific production emerged.
- The term “bibliometrics” was coined by Paul Otlet (1934) and consolidated by Alan Pritchard (1969).
- Advances influenced its development in documentation, statistics, and information technology.
- The first studies on author productivity and article distribution were published in the 1920s and 1940s.
- Lotka (1926) formulated the Law of Scientific Productivity, which describes the unequal frequency with which researchers publish.
- Bradford (1934) proposed his Law of Dispersion, which explains how relevant articles are concentrated in a small number of core journals.
- Zipf (1949) introduced the Law of Word Frequency, which serves as the basis for the analysis of terms and co-occurrences.
- In the 1950s and 1960s, Eugene Garfield founded the Institute for Scientific Information (ISI) and developed the Science Citation Index (SCI).
- The SCI revolutionized the measurement of science by enabling the tracking of citation networks between publications.
- In the 1970s and 1980s, the field became institutionalized with the emergence of the journal Scientometrics and the first international conferences.
- At this stage, bibliometrics expanded into scientometrics, focusing on the dynamics and politics of science.
- Starting in the 1990s, the use of electronic databases and specialized software allowed for broader and more accurate analysis.

- The emergence of the Internet and search engines radically transformed access to and the collection of scientific information.
- With the arrival of Google Scholar (2004) and Scopus (2004), bibliometric sources and impact indicators diversified.
- Since 2010, altmetrics and webmetrics have emerged, focusing on the social and digital impact of science.
- In this modern stage, bibliometrics is integrated with artificial intelligence, big data, and open science.
- Current methods enable mapping collaboration networks, thematic trends, and cognitive structures.
- The historical development reflects a transition from a simple quantitative approach to a multidimensional and ethical view of scientific evaluation.
- Contemporary bibliometrics combines the tradition of Garfield, Lotka, and Bradford with advanced digital tools.
- In short, its evolution has made bibliometrics an essential pillar of research, knowledge management, and science policy.

Self-assessment questions

1. Who coined the term "bibliometrics" and in what year did its use become established?
2. What does Lotka's Law describe in relation to scientific productivity?
3. What is the central principle of Bradford's Law?
4. What did Zipf contribute to the development of bibliometrics?
5. How important was Eugene Garfield in the history of the discipline?
6. What role did the creation of the Science Citation Index play in the 1960s?
7. Why did the 1970s and 1980s mark the institutionalization of bibliometrics?
8. How did electronic databases and the Internet influence the evolution of the field?
9. What contributions did altmetrics introduce after 2010?
10. How does contemporary bibliometrics differ from its early stages?

BIBLIOGRAPHY

1.Sugimoto CR, Larivière V. Measuring research: What everyone needs to know. Oxford: Oxford University Press; 2018. https://global.oup.com/academic/product/measuring-research-9780190640125

2. Glänzel W. Bibliometrics as a research field: A course on theory and application of bibliometric indicators. Leuven: KU Leuven; 2003. https://www.kuleuven.be/metaforum/docs/pdf/lecture_glanzel_bibliometrics.pdf

3. Waltman L. A review of the literature on citation impact indicators. J Informetrics. 2016;10(2):365–91. doi:10.1016/j.joi.2016.02.007

4. Cronin B, Sugimoto CR, editors. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. Cambridge (MA): MIT Press; 2014. doi:10.7551/mitpress/9780262026792.001.0001

5. Hood WW, Wilson CS. The literature of bibliometrics, scientometrics, and informetrics. Scientometrics. 2001;52(2):291–314. doi:10.1023/A:1017919924342

BIBLIOGRAPHIC REFERENCES

1. Garfield E. Citation indexes for science. Science. 1955;122(3159):108–11.

2. Garfield E. “Science Citation Index”—A new dimension in indexing. Science. 1964;144(3619):649–54.

3. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: a multidisciplinary comparison of coverage via citations. Scientometrics. 2021;126(1):871–906. https://doi.org/10.1007/s11192-020-03690-4

4. Garfield E. The evolution of the science citation index. International Microbiology. 2007;10(1):65–9.

5. Lotka AJ. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 1926;16(12):317–23.

6. Bradford SC. Classic paper: Sources of information on specific subjects. Collection Management. 1976;1(3–4):95–103. https://doi.org/10.1300/J105v01n03_06

7. Patsopoulos NA, Analatos AA, Ioannidis JPA. Relative citation impact of various study designs in the health sciences. JAMA. 2005;293(19):2362–6. https://doi.org/10.1001/jama.293.19.2362

8. Agencia Nacional de Investigación y Desarrollo (ANID). Concurso de Proyectos Fondecyt Regular 2024. 2024. https://anid.cl/concursos/concurso-de-proyectos-fondecyt-regular-2024/

Part II / Parte II

DATA PREPARATION

PREPARACIÓN DE DATOS

AG
EDITOR

Chapter 3 / Capítulo 3

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch03

Tools for Bibliometric Analysis/ Herramientas para el análisis bibliométrico

3.1. Introduction to bibliometric tools

Bibliometric software represents the methodological backbone that transforms raw data into actionable knowledge. Without these specialized tools, researchers would face the practical impossibility of manually analyzing the massive volumes of scientific publications that characterize the contemporary era. Automation allows thousands of references to be processed in minutes, identifying patterns of collaboration, thematic trends, and citation networks that would otherwise remain hidden in the noise of information. This computational capacity not only optimizes time resources but also fundamentally expands the frontiers of what can be analyzed, enabling research questions that were previously unattainable due to technical and operational limitations to be addressed.

The choice of the right software directly determines the validity and depth of bibliometric findings. Each tool incorporates particular methodological assumptions that condition the types of analysis possible and the interpretations derived. Platforms such as VOSviewer prioritize network visualization, while Bibliometrix emphasizes statistical indicators, and Python offers algorithmic flexibility. This specialization creates a complementary ecosystem where the strategic combination of tools overcomes individual limitations. Competence with multiple software programs thus becomes a fundamental skill for the contemporary bibliometrist seeking to produce robust, multidimensional scientific evaluations.

The evolution of bibliometric software reflects and simultaneously drives the theoretical development of the discipline. Current tools incorporate advances such as natural language processing for content analysis or machine learning algorithms for detecting emerging topics. This symbiosis between technological development and methodological sophistication has transformed bibliometrics from a basic descriptive exercise into a predictive analytical science. Modern researchers must understand both the bibliometric fundamentals and the technical capabilities of the software to design studies that leverage the full analytical potential of the contemporary digital ecosystem.

This book does not aim to provide an in-depth look at the tools used to create bibliometrics, but rather to equip the reader with the minimum knowledge needed to use them, along with guidance for those who wish to explore them in greater depth.

3.1.1. Classification of tools and selection criteria

The current bibliometric software ecosystem is highly diverse, responding to different methodological needs and levels of expertise. This variety can be organized into clearly defined categories according to technical complexity and specific analytical objectives. Understanding this fundamental taxonomy is the first step toward making an appropriate selection that optimizes available research resources.[(1)]

At the base of the tool pyramid is basic analysis software, designed for users who require quick calculations of fundamental indicators. These solutions prioritize accessibility over analytical depth, offering intuitive interfaces that minimize the learning curve. Their main advantage lies in their ability to deliver immediate results without specialized technical knowledge, though their methodological flexibility is limited for complex research.

Comprehensive analysis platforms represent the next level of technological sophistication, integrating multiple functionalities into unified environments. Often developed as complete suites, they allow complete bibliometric workflows to be executed without changing environments. Their modular architecture supports everything from fundamental descriptive analysis to advanced text mining and thematic pattern detection across large volumes of scientific literature.

Specialized visualization tools occupy a particular space within the ecosystem, focusing on the graphical representation of networks and knowledge structures. Their added value lies in specific algorithms for network layout and cartographic techniques that transform bibliometric data into interpretable maps. These solutions complement rather than replace the quantitative analysis performed on other platforms.

At the pinnacle of technical complexity are solutions that can be customized through programming, which give total control over every methodological aspect. This code-based approach sacrifices immediacy for absolute flexibility, allowing the implementation of novel techniques not available in commercial tools. It is the preferred option for cutting-edge bibliometric research that requires original methodological developments.

The selection between these categories should be guided by specific criteria that transcend personal preferences. Data volume is a primary consideration, as tools designed for meta-analysis of thousands of records differ significantly from those optimized for focused bibliographic studies. The software's scalability determines its suitability for projects of varying scope.

Available computational resources represent another decisive factor that is often underestimated in research planning. Some advanced visualization tools require substantial graphics processing capabilities, while web-based solutions externalize these technical requirements. A realistic assessment of available technological infrastructure prevents bottlenecks during the analytical phase.

Interoperability with specific data sources deserves special consideration, given the current fragmentation of the bibliographic ecosystem. Native compatibility with Scopus, WoS, Dimensions, or PubMed directly conditions the technical feasibility of many projects. Solutions that offer multiple, standardized connectors significantly reduce preprocessing and normalization efforts.

The learning curve for each tool must be weighed against the time available and the research team's expertise. While some modern platforms prioritize user experience through intuitive graphical interfaces, others require specialized knowledge that can slow down the initial stages of the project. Existing training and training capacity are critical variables in this equation.

Reproducibility and methodological transparency emerge as decisive criteria for high-impact research. Tools that allow the export of complete workflows and executable scripts facilitate peer review and the replicability of studies. This feature is particularly valuable in scientific evaluation contexts where auditability is a fundamental requirement.

3.2. Preparing the work environment for R

Some of the giants of bibliometric research are the R Bibliometrix[1] and Python PyBibx[2] libraries, respectively. Both languages require prior configuration and installation to be used. These will be addressed below.

3.2.1. Installation and configuration of R and RStudio

Preparing the R environment to run code in R begins with downloading the latest version of R from the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/), then selecting the installer for the operating system. After completing the basic installation, download RStudio Desktop (https://posit.co/products/open-source/rstudio), the integrated development environment that optimizes R workflows. The initial configuration includes defining the working directory, adjusting appearance preferences, and setting performance options based on the computer's capabilities. These are selected as they are displayed in the installer's options. This solid foundation ensures a stable environment for running bibliometric packages.

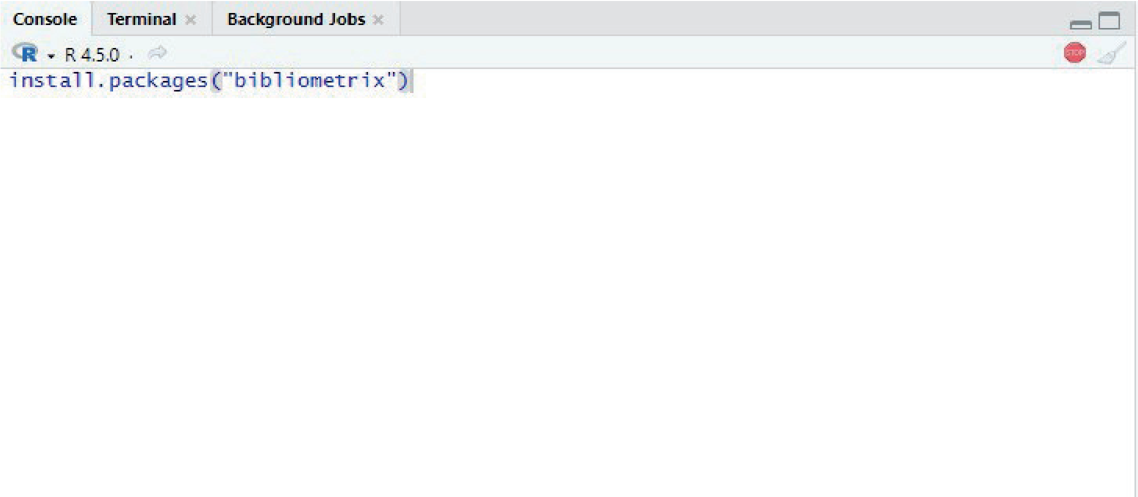
A fundamental concept to understand when working with these programming languages is that of a library or package, a collection of pre-written code that extends the language's basic capabilities. They function as a set of specialized tools, providing functions and methods for specific tasks without requiring developers to create them from scratch. For example, a data science library may contain algorithms for statistical analysis or graphics. This not only greatly speeds up the development process but also ensures reliability and efficiency by leveraging code that the community has already tested. In essence, they are fundamental components that allow complex applications to be built in a modular and efficient manner.

Package management in R is a crucial step for bibliometric analysis. It is recommended to start by installing essential packages, such as Bibliometrix, bibliometR, and the tidyverse, from the official CRAN repository. This process will be discussed in more detail later. Configuring personal libraries avoids version conflicts and facilitates project portability. This meticulous preparation prevents errors during subsequent analytical processes.

3.2.2. Specialized tools in R

Bibliometrix: Complete installation and configuration

Bibliometrix is installed in RStudio using the command install.packages("bibliometrix"). After installation, loading the library with the library(Bibliometrix) enables all its features for comprehensive bibliometric analysis.



```
install.packages("bibliometrix")
```

Figure 3.1. Installation of R Bibliometrix

The Bibliometrix package offers three different interfaces to suit different user profiles.

The biblioshiny() function launches a web-based graphical interface ideal for beginners or quick exploratory analyses. For intermediate users, basic functions such as convert2df() and biblioAnalysis() provide direct control over the workflow. Advanced users can access all functions via direct programming in R, enabling sophisticated methodological customization. This flexibility makes Bibliometrix a complete solution for all types of bibliometric projects. In this text, the last two options will be used for greater control over the generated content.

3.2.3. Main features and fundamental analysis

Bibliometrix deploys a comprehensive analytical repertoire that begins with fundamental descriptive analysis of scientific output. The biblioAnalysis function automatically generates fundamental indicators, including temporal growth, most productive authors, and leading journals. The identification of collaborations using networkPlot reveals co-authorship patterns at the individual, institutional, and national levels. These descriptive functionalities provide the necessary contextual basis before undertaking more specialized analyses such as conceptual mapping or citation studies.

Content analysis and thematic trends are among Bibliometrix’s most powerful capabilities. The conceptual structure function uses co-word analysis to identify thematic clusters and their evolution over time. Co-citation network mapping using cocMatrix reveals the intellectual structure of the field of study. The thematicMap function visualizes topics according to their degree of development and relevance, distinguishing between emerging, basic, niche, and driving issues. These tools facilitate the identification of knowledge frontiers and research opportunities. How to use and interpret them will be discussed in more detail in later chapters.

The R ecosystem for bibliometrics is enriched with specialized libraries that complement Bibliometrix. One of these is bibliometR, which offers additional functionalities for citation network analysis and community detection. RefManageR facilitates advanced management of bibliographic references in different citation styles. There are others, but currently the most widely used is RBibliometric, which will be the focus of this book.

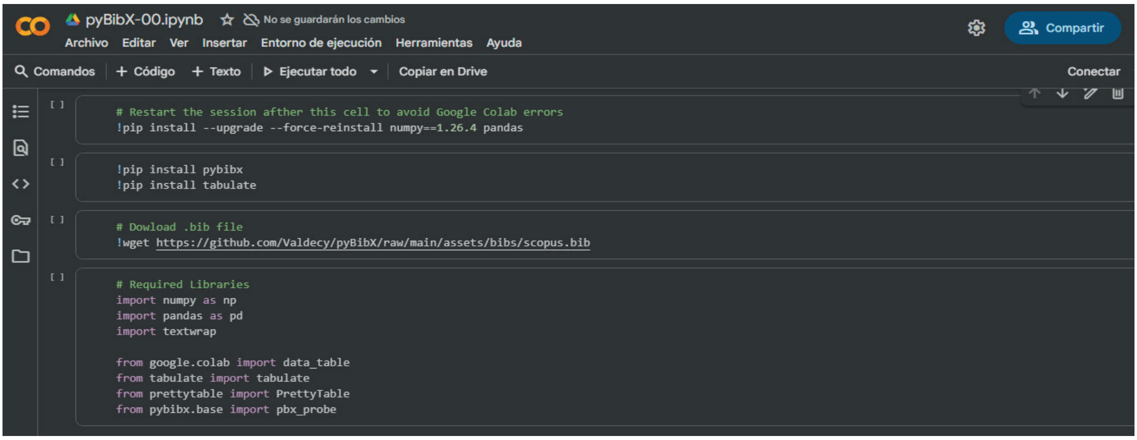
3.3. Working with Google Colab for PyBibX

Google Colab is the ideal platform for implementing PyBibX, eliminating the need for local configuration. Free access via a Google account immediately provides a preconfigured Python environment with the main scientific libraries. Connection to enhanced computational resources, including GPUs and TPUs, significantly speeds up the processing of large bibliometric datasets. This cloud infrastructure ensures consistency among collaborators and simplifies the reproducibility of analyses.

Since Google Colab runs in the browser, no prior equipment configuration is necessary; only internet access from the device’s browser is required. It will only be essential to install the required libraries, as in the case of RBibliometric. This is done with the *pip install pybibx command.*

The official PyBibX documentation (https://pypi.org/project/pybibx/) provides a list of codes for the types of files to be imported for analysis, so it is not necessary to memorize each code or write them manually. Open the link according to the file and start working.

There are other alternatives to Colab, such as Anaconda. Still, in this text, Colab will be used because its session management optimizes work with PyBibX in extensive bibliometric projects, the connection to runtime with high RAM prevents failures during the processing of large document corpora, and the scheduling of periodic executions using cron jobs allows analyses to be kept up to date without constant manual intervention. Automatic export of results to Google Drive ensures the persistence of findings beyond the active session on the platform.


pyBibX-00.ipynb
No se guardarán los cambios
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda
Compartir
Comandos
+ Código
+ Texto
Ejecutar todo
Copiar en Drive
Conectar

```
# Restart the session afther this cell to avoid Google Colab errors
!pip install --upgrade --force-reinstall numpy==1.26.4 pandas
```

```
!pip install pybibx
!pip install tabulate
```

```
# Dowload .bib file
!wget https://github.com/Valdecy/pyBibX/raw/main/assets/bibs/scopus.bib
```

```
# Required Libraries
import numpy as np
import pandas as pd
import textwrap

from google.colab import data_table
from tabulate import tabulate
from prettytable import PrettyTable
from pybibx.base import pbx_probe
```

Figure 3.2. PyBibx project in Google Colab

In both R Bibliometrix and Pybibx, after importing the libraries, it is necessary to load the data file (s) for the study. Each of these libraries has its own particular way of importing and subsequently analyzing them.

R Bibliometrix

In R (Bibliometrix), the convert2df() function reads an exported file from databases such as Scopus or WoS and converts it into a data frame that the library can work with.

```
# Load the library
library(bibliometrix)

# Import and convert the file downloaded from Scopus/WoS into a dataset called my_dataset
my_dataset <- convert2df(“my_scopus_file.bib”, dbsource = “scopus”, format = “bibtex”)
```

Note: anything after # is a comment; even if you paste it into RStudio or Colab, it will not be executed as code.

Pybibx

In Python (Pybibx) with Google Colab, you must first upload the file to the environment and then use the load function of the dataset module to load it into a Collection object, which is the central data structure. This may require other library imports. The official Pybibx website provides a list of code examples for each type of data file, for example, if they were exported as .bib from Scopus, or txt from PubMed, among others, depending on their source.

```
# Install and import the library (in Google Colab)
!pip install pybibx
from pybibx import dataset

# Upload the .bib or .csv file to Colab
from Google. colab import files
```

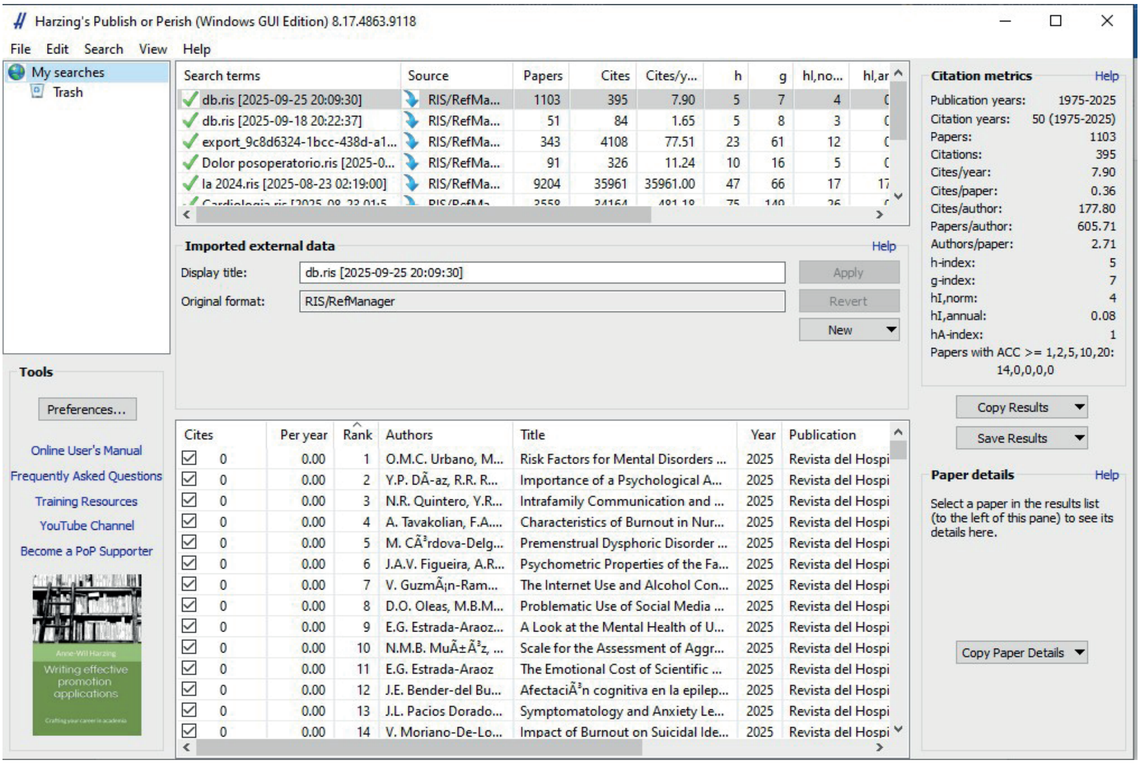
```
uploaded = files.upload()

# Load the dataset into a Collection called my_dataset
my_dataset = dataset.load(list(uploaded.keys())[0])
```

You don't need to know how to program to use this book; the code shown can be copied as is and will work.

3.4. Publish or Perish: Analysis with Google Scholar

3.4.1. Installation, configuration, and search types



Search terms	Source	Papers	Cites	Cites/y...	h	g	hI,no...	hI,ar
✓ db.ris [2025-09-25 20:09:30]	RIS/RefMa...	1103	395	7.90	5	7	4	0
✓ db.ris [2025-09-18 20:22:37]	RIS/RefMa...	51	84	1.65	5	8	3	0
✓ export_9c8d6324-1bcc-438d-a1...	RIS/RefMa...	343	4108	77.51	23	61	12	0
✓ Dolor posoperatorio.ris [2025-0...	RIS/RefMa...	91	326	11.24	10	16	5	0
✓ la 2024.ris [2025-08-23 02:19:00]	RIS/RefMa...	9204	35961	35961.00	47	66	17	17
✓ Cardiologia.ris [2025-08-23 01:5	RIS/RefMa...	3558	34164	481.18	75	140	26	

Imported external data

Display title: db.ris [2025-09-25 20:09:30]

Original format: RIS/RefManager

Citation metrics

Publication years:	1975-2025
Citation years:	50 (1975-2025)
Papers:	1103
Citations:	395
Cites/year:	7.90
Cites/paper:	0.36
Cites/author:	177.80
Papers/author:	605.71
Authors/paper:	2.71
h-index:	5
g-index:	7
hI,norm:	4
hI,annual:	0.08
hA-index:	1
Papers with ACC >= 1,2,5,10,20:	14,0,0,0,0

Paper details

Select a paper in the results list (to the left of this pane) to see its details here.

Cites	Per year	Rank	Authors	Title	Year	Publication
☑ 0	0.00	1	O.M.C. Urbano, M...	Risk Factors for Mental Disorders ...	2025	Revista del Hospi
☑ 0	0.00	2	Y.P. DÃ-az, R.R. R...	Importance of a Psychological A...	2025	Revista del Hospi
☑ 0	0.00	3	N.R. Quintero, Y.R...	Intrafamily Communication and ...	2025	Revista del Hospi
☑ 0	0.00	4	A. Tavakolian, F.A....	Characteristics of Burnout in Nur...	2025	Revista del Hospi
☑ 0	0.00	5	M. CÃ³rdova-Delg...	Premenstrual Dysphoric Disorder ...	2025	Revista del Hospi
☑ 0	0.00	6	J.A.V. Figueira, A.R...	Psychometric Properties of the Fa...	2025	Revista del Hospi
☑ 0	0.00	7	V. GuzmÃ¡n-Ram...	The Internet Use and Alcohol Con...	2025	Revista del Hospi
☑ 0	0.00	8	D.O. Oleas, M.B.M...	Problematic Use of Social Media ...	2025	Revista del Hospi
☑ 0	0.00	9	E.G. Estrada-Araoz...	A Look at the Mental Health of U...	2025	Revista del Hospi
☑ 0	0.00	10	N.M.B. MuÃ±Ã³z, ...	Scale for the Assessment of Aggr...	2025	Revista del Hospi
☑ 0	0.00	11	E.G. Estrada-Araoz	The Emotional Cost of Scientific ...	2025	Revista del Hospi
☑ 0	0.00	12	J.E. Bender-del Bu...	AfectaciÃ³n cognitiva en la epilep...	2025	Revista del Hospi
☑ 0	0.00	13	J.L. Pacios Dorado...	Symptomatology and Anxiety Le...	2025	Revista del Hospi
☑ 0	0.00	14	V. Moriano-De-Lo...	Impact of Burnout on Suicidal Ide...	2025	Revista del Hospi

Figure 3.3. Harizing's Publish or Perish home screen

The installation of Publish or Perish begins with a free download from the official Harzing.com website (https://harzing.com/), selecting the version compatible with the user's operating system. The installation process is simple and does not require advanced technical configuration, though it is essential to allow network connections so the software can access Google Scholar data. After installation, the main interface displays clearly organized options that let you start performing bibliometric searches immediately after initial configuration.

Optimal software configuration involves customizing preferences to specific research needs. Among the most relevant settings is the selection of the primary search engine, where Google Scholar is the most commonly used default option. It is crucial to configure the wait time between queries to avoid being blocked by Google, setting intervals that simulate human behavior. Customizing export formats and result limits completes the preparation for efficient and sustainable use of the tool.

The search types available in Publish or Perish are tailored to different bibliometric objectives. The author search allows you to locate and analyze the impact of specific researchers using different variations of their names. The article search makes it easy to track citations of individual publications by combining title, author, and year. The general search by research terms provides a broad overview of the literature on particular topics, while the journal search evaluates the impact of specific periodicals.

Each search type requires specific strategies to optimize data retrieval. For authors, it is essential to try different combinations of names and initials to counteract Google Scholar's limitations in disambiguation. In thematic searches, the strategic combination of Boolean operators and exact phrases significantly improves the accuracy of results. Setting filters by publication year, language, and document type refines result sets before analysis, saving time in subsequent processing of retrieved information.

The software offers advanced search capabilities that overcome the limitations of the Google Scholar web interface. The ability to search in multiple languages extends the scope of bibliographic retrieval beyond Anglo-centric literature. The search function by institutional affiliation, although limited by Google Scholar's coverage, provides valuable insights into the productivity of different research centers. These features make Publish or Perish an indispensable tool for bibliometric analysis, complementary to commercial databases.

The effectiveness of searches depends largely on understanding Google Scholar's inherent limitations. Uneven coverage across disciplines, the inclusion of non-peer-reviewed documents, and duplicate records require careful interpretation of results. The software partially mitigates these problems through cleaning and deduplication algorithms, but the researcher's critical judgment remains essential to validate the quality of the data obtained before formal bibliometric analysis.

Publish or Perish significantly expands its capabilities by integrating direct connections to Scopus, Web of Science, PubMed, and other databases using valid institutional credentials. This functionality turns the tool into a bridge between Google Scholar metrics and the databases above, enabling unique comparative analyses. Users can authenticate themselves through institutional subscriptions to access standardized data of higher bibliographic quality. This integration allows for cross-referencing across sources, identifying discrepancies and complementarities in citation coverage for a more comprehensive assessment of research impact.

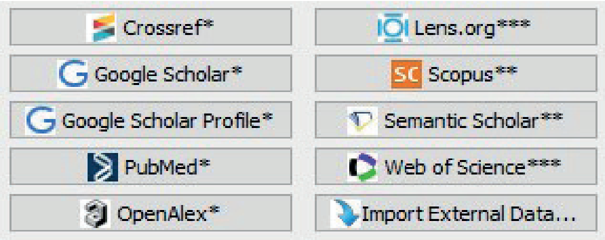


Figure 3.4. Publish or Perish data import block

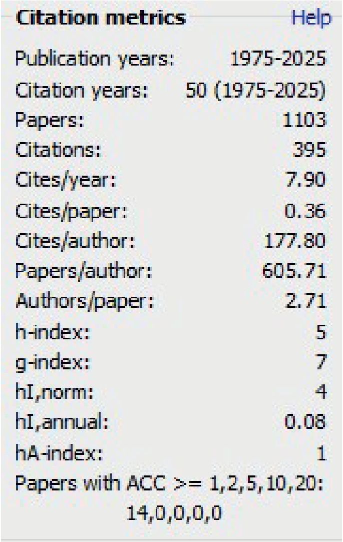
The ability to import data from multiple sources is a distinctive advantage of Publish or Perish over other bibliometric tools. The software accepts standard formats such as RIS, BibTeX, and CSV from reference managers or exports from other platforms. This interoperability facilitates the consolidation of scattered bibliographic information into a single metric analysis environment. Researchers can combine datasets from different sources to create customized collections that fully reflect their scientific output, thereby overcoming the coverage limitations of individual sources.

The import process allows basic metadata to be enriched with impact indicators calculated by Publish or Perish. When loading references from Zotero, Mendeley, or Scopus exports, the software automatically searches for corresponding citations in Google Scholar and other available sources. This functionality is particularly valuable for updating existing bibliographic collections with recent citation data without manual searches. The result is an integrated dataset that preserves the original metadata while adding the most up-to-date impact metrics.

3.4.2. Data extraction and metrics provided

The Publish or Perish system performs comprehensive searches across Google Scholar, Scopus, and Web of Science, capturing essential metadata, including titles, authors, affiliations, publication years, and sources. The tool automatically processes citation networks to calculate advanced indicators, identifying both citations received and temporal impact trends. This comprehensive process ensures that complete datasets are obtained for rigorous bibliometric analysis.

The primary metrics automatically calculated include the h-index, g-index, and individual h-index, providing different perspectives on research impact. The software also generates the total number of citations, citations per year, citations per article, and average citations per work. These indicators are complemented by temporal statistics that reveal patterns of productivity and impact evolution throughout the research career. Each metric is presented with its detailed calculation to facilitate contextual interpretation.



Citation metrics Help

Publication years:	1975-2025
Citation years:	50 (1975-2025)
Papers:	1103
Citations:	395
Cites/year:	7.90
Cites/paper:	0.36
Cites/author:	177.80
Papers/author:	605.71
Authors/paper:	2.71
h-index:	5
g-index:	7
hI,norm:	4
hI,annual:	0.08
hA-index:	1

Papers with ACC >= 1,2,5,10,20:
14,0,0,0,0

Figure 3.5. Publish or Perish analysis results block

The tool offers sophisticated comparative analyses that allow researchers, institutions, or journals to be evaluated against disciplinary benchmarks. The system calculates impact percentiles, relative positions, and comparative trends using standardized algorithms. This functionality is particularly valuable for bibliometric studies that require contextualizing performance within specific fields of knowledge. The results include clear visualizations that facilitate the identification of strengths and areas for improvement in the analyzed research profiles.

The extracted data includes qualitative information essential to correctly interpreting quantitative metrics. The software captures full titles, abstracts when available, and the most relevant cited references, allowing for an understanding of the reasons behind the measured impact, differentiating between citations of recognition, methodological citations, and academic controversies. The integration of qualitative and quantitative dimensions substantially enriches the resulting bibliometric analysis.

3.4.3. Export and integration with other tools

Publish or Perish offers advanced export capabilities that facilitate interoperability with the bibliometric ecosystem. Users can export complete results in CSV, RIS, or BibTeX formats, preserving all metadata and calculated metrics. This flexibility allows data to be integrated with specialized tools such as VOSviewer for network visualization, Bibliometrix for advanced statistical analysis, or CitNetExplorer for studying citation networks. The export maintains the relational structure of the data, ensuring that connections between authors, articles, and citations remain intact during transfer.

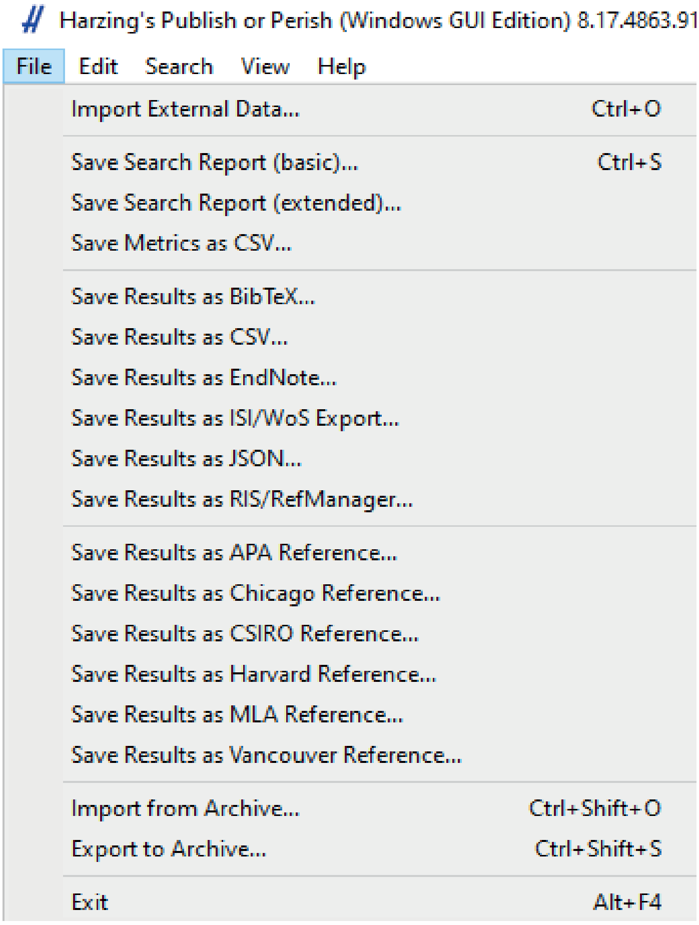

Harzing's Publish or Perish (Windows GUI Edition) 8.17.4863.91
File
Edit
Search
View
Help
Import External Data... Ctrl+O
Save Search Report (basic)... Ctrl+S
Save Search Report (extended)...
Save Metrics as CSV...
Save Results as BibTeX...
Save Results as CSV...
Save Results as EndNote...
Save Results as ISI/WoS Export...
Save Results as JSON...
Save Results as RIS/RefManager...
Save Results as APA Reference...
Save Results as Chicago Reference...
Save Results as CSIRO Reference...
Save Results as Harvard Reference...
Save Results as MLA Reference...
Save Results as Vancouver Reference...
Import from Archive... Ctrl+Shift+O
Export to Archive... Ctrl+Shift+S
Exit Alt+F4

Figure 3.6. Publish or Perish export menu

Integration with bibliographic managers is a particularly valuable feature for researchers. Data exported in RIS format can be imported directly into Zotero, Mendeley, or EndNote, enriching personal libraries with calculated impact metrics. This feature eliminates the need for manual updating processes and ensures that bibliographic collections maintain up-to-date metric information. Bidirectional synchronization allows simultaneous work on reference management and bibliometric analysis within cohesive workflows.

For advanced quantitative analysis, exporting in CSV format provides immediate compatibility with statistical software such as R, Python, or SPSS. The resulting tables include standardized columns for each bibliometric variable, facilitating processing using custom scripts. Researchers can combine this data with additional information from other sources to create enriched datasets for multivariate models. This capability transforms Publish or Perish into an initial extraction tool within complex analytical pipelines.

Export options include customizable settings to suit specific research requirements. Users can select subsets of metrics, filter by time ranges, or choose particular encoding formats. This granularity ensures that data transferred to other tools contains exactly the information needed for each type of subsequent analysis. The combination of flexibility and precision in exporting positions in Publish or Perish is a fundamental component of integrated bibliometric workflows.

3.4.4. Limitations and best practices for use

Publish or Perish has significant limitations that users must recognize to interpret the results correctly. Its primary reliance on Google Scholar introduces inherent biases, including uneven coverage across disciplines, the inclusion of non-peer-reviewed documents, and record duplication. The tool lacks robust mechanisms for automatic author disambiguation, which can distort individual profiles when multiple researchers share the same name. Furthermore, the calculated metrics primarily reflect quantitative impact, without accounting for qualitative dimensions such as journal prestige or the type of scientific contribution.

Essential best practices include systematic cross-checking of results with complementary databases such as Scopus or Web of Science. Users should implement conservative search strategies, using multiple variations of author names and manually validating the most relevant profiles. It is crucial to contextualize metrics within specific disciplinary norms, recognizing that identical h-index values have different meanings in theoretical physics and philosophy. Methodological transparency requires documenting all search and data cleaning decisions in final reports.

Responsible interpretation of metrics requires understanding their methodological foundations and technical limitations. Users should complement quantitative metrics with qualitative assessment, reviewing the most-cited works to determine their actual contribution to the field. This balanced approach prevents erroneous conclusions drawn solely from numerical indicators without interpretive context.

The ethical management of the data obtained involves respecting the use limits established by the sources and avoiding overloading their servers. Massive searches should be scheduled with reasonable intervals between queries, and results should be stored in accordance with standard security protocols. Institutional users should establish clear guidelines for the appropriate use of metrics in academic evaluations, aligning with the principles of the DORA Declaration to avoid reductionism in research assessment.

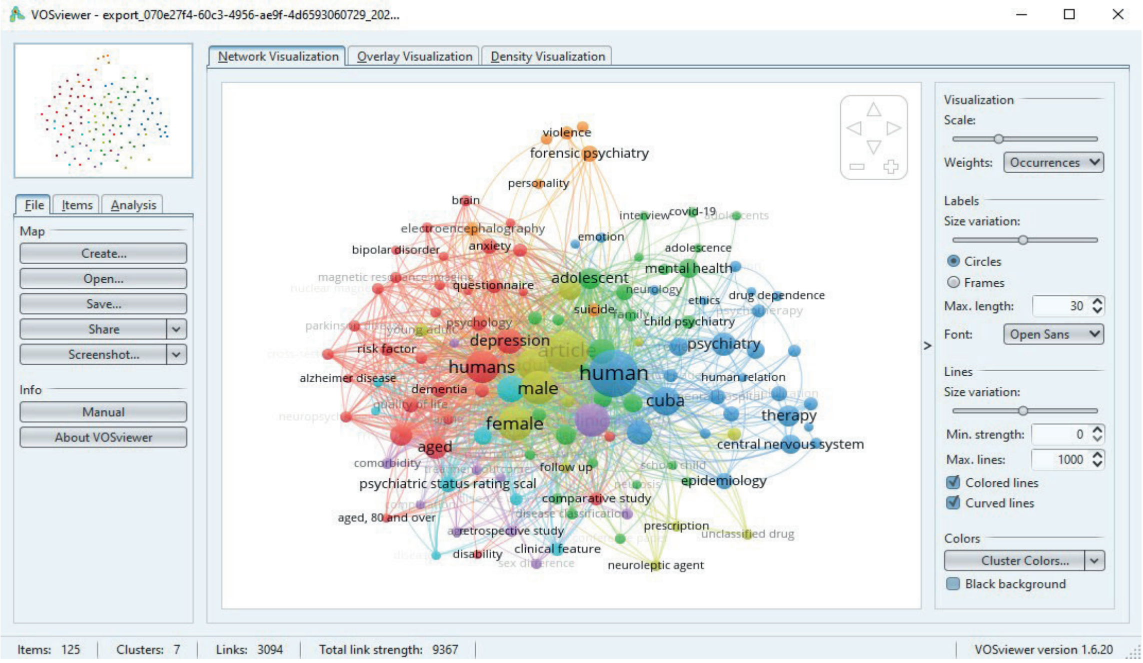
3.5. Advanced visualization tools

The visualization of bibliometric data is a crucial stage in interpreting and communicating results, transforming complex datasets into intuitive graphical representations that reveal underlying patterns, relationships, and trends in the scientific literature. These specialized tools allow researchers and evaluators to identify collaboration networks, thematic structures, and knowledge dynamics that would remain hidden in conventional tables and lists, facilitating both specialized analysis and the effective dissemination of bibliometric findings to multidisciplinary audiences.

3.5.1. VOSviewer: installation and creation of scientific maps

The installation of VOSviewer begins with a free download from the official website of the Center for Science and Technology Studies at Leiden University, available for Windows, Mac OS, and Linux (https://www.VOSviewer.com/download). The installation process is simple and does not require advanced technical configuration. However, it is advisable to verify that the system has the latest version of Java to ensure optimal performance of all features. Once installed, the work environment is organized into logical modules that guide the user through the entire process of creating scientific maps.

Map creation in VOSviewer begins with the import of data from bibliographic sources such as Scopus, Web of Science, or PubMed, which have been previously processed into compatible formats. The next chapter will describe the process of obtaining this data. The software offers three main types of analysis: co-authorship networks to visualize scientific collaborations, co-citation networks to reveal the intellectual structure of a field, and term co-occurrence maps to identify research topics. Each type of analysis uses specific layout and clustering algorithms that optimize the visual representation of bibliometric relationships.



VOSviewer - export_070e27f4-60c3-4956-ae9f-4d6593060729_202...

Visualization
Scale:
Weights: Occurrences
Labels
Size variation:
● Circles
○ Frames
Max. length: 30
Font: Open Sans
Lines
Size variation:
Min. strength: 0
Max. lines: 1000
☑ Colored lines
☑ Curved lines
Colors
Cluster Colors...
□ Black background

Items: 125 | Clusters: 7 | Links: 3094 | Total link strength: 9367 | VOSviewer version 1.6.20

Figure 3.7. Main view of VOSviewer, with a project displayed

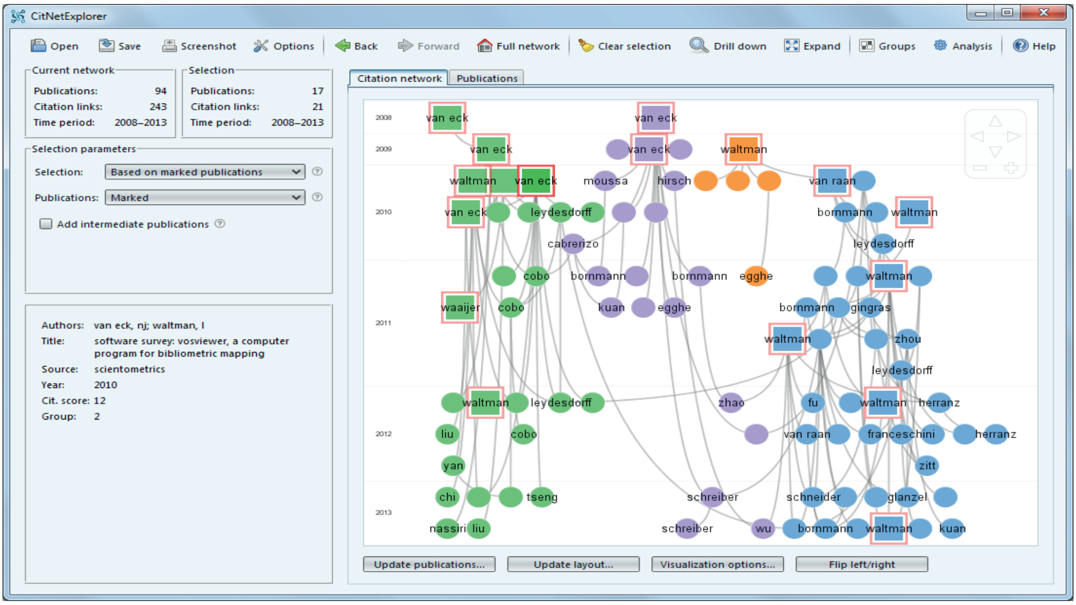
The mapping process involves several configuration stages in which the user defines key parameters, such as the minimum occurrence threshold, normalization method, and clustering algorithm. VOSviewer applies advanced dimensionality reduction techniques using the VOS (Visualization of Similarities) algorithm, which positions elements in a two-dimensional space while preserving their similarity relationships. The tool automatically generates color-coded clusters representing thematic communities or collaboration groups, facilitating the visual identification of structural patterns.

Map customization allows adjustment of multiple visual aspects, such as node sizes, link thicknesses, text fonts, and color schemes. Nodes can be sized according to different metrics, such as frequency of occurrence, number of citations, or centrality in the network, while link thickness reflects the strength of the relationships between elements. These customization options allow the visualization to be adapted to the specific needs of each analysis and target audience, improving the communicative clarity of the results.

Interpreting the generated maps requires understanding both the relative position of the elements and their grouping into thematic clusters. The distance between nodes indicates their degree of thematic or collaborative relationship, while the spatial distribution reveals the overall structure of the field of study. The maps allow for the identification of emerging themes, central authors, leading institutions, and patterns of international collaboration, providing valuable insights for scientific evaluation and future research planning.

VOSviewer includes advanced features such as temporal analysis using overlay maps, export to vector formats for academic publications, and zoom and filtering tools to explore specific areas of interest. Integration with other bibliometric software using standard exchange formats completes a robust visualization ecosystem that supports bibliometric research from initial exploratory analysis to the final presentation of complex results.

3.5.2. CitNetExplorer: citation network analysis



Current network

Publications:	94
Citation links:	243
Time period:	2008–2013

Selection

Publications:	17
Citation links:	21
Time period:	2008–2013

Selection parameters

Selection: Based on marked publications

Publications: Marked

□ Add intermediate publications

Authors: van eck, nj; waltman, l

Title: software survey: vosviewer, a computer program for bibliometric mapping

Source: scientometrics

Year: 2010

Cit. score: 12

Group: 2

Source: CitNetExplorer official website (https://www.citnetexplorer.nl).

Figure 3.8. Main view of CitNetExplorer, with project graph

CitNetExplorer specializes in the analysis and visualization of citation networks, allowing users to explore the structure and evolution of scientific fields through the relationships between publications. Developed by Leiden University, this software enables users to identify key publications, trace the evolution of research lines, and analyze citation patterns over time. Its unique approach facilitates the study of intellectual heritage and knowledge trajectories within specific scientific disciplines.

The software is easy to install: download it from the official website and decompress the ZIP file into the desired directory (https://www.citnetexplorer.nl/download). CitNetExplorer does not require conventional installation; it runs directly from the JAR file with Java Runtime Environment 8 or higher. This portability allows the tool to be used across different systems without complex configuration, though it is recommended to verify execution permissions on Unix-based systems to ensure proper operation.

The analysis process begins with importing data from Web of Science or Scopus using specific export formats that preserve the cited references. The tool automatically constructs the citation network, where nodes represent publications and links represent citation relationships. Visualization algorithms organize publications chronologically, clearly showing the field's evolution and enabling the identification of seminal works that serve as connecting points between different lines of research.

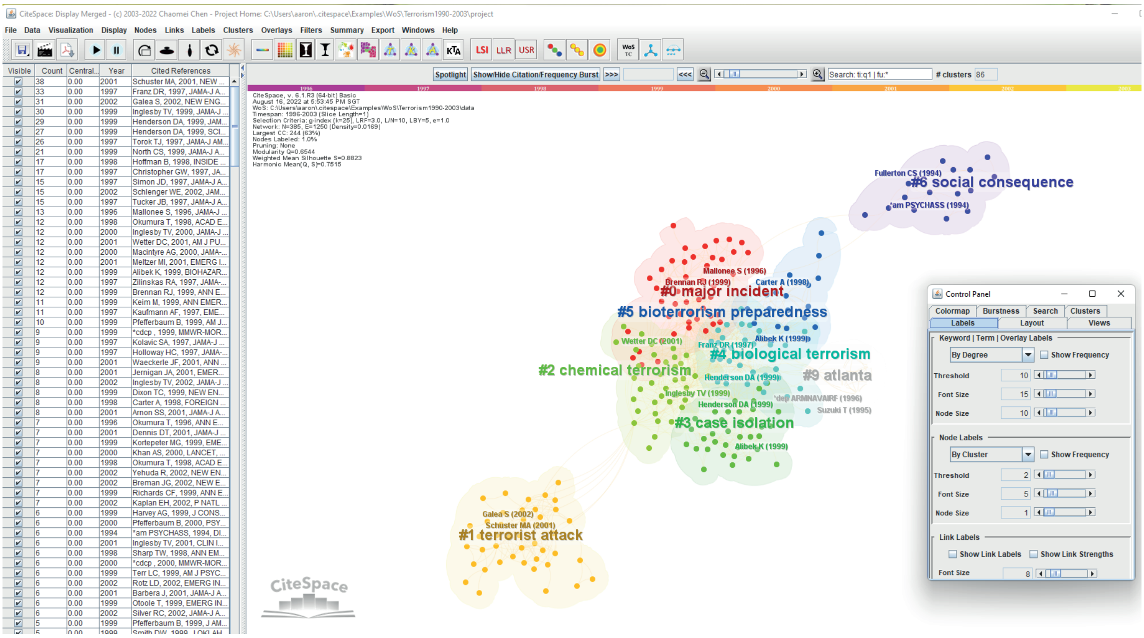
Exploration features include zoom and time-filtering tools that let users focus on specific periods. Users can select key publications to visualize their local citation network and identify predecessor and successor works. Citation path analysis reveals how scientific ideas are transmitted and transformed through different publications, providing insights into the dynamics of knowledge dissemination within the academic community.

The identification of thematic clusters is performed using algorithms that detect highly connected communities of publications. These groups represent specific subfields or lines of research, whose evolution can be tracked over time. CitNetExplorer allows you to analyze how different intellectual traditions emerge, converge, or diverge, offering unique insights into the structuring of scientific knowledge that complement analyses from other bibliometric visualization tools.

Export options include detailed reports with network metrics, high-resolution images for publications, and processed data for complementary analysis. The tool allows users to save complete work sessions, facilitating continuity in extensive research. This ability to preserve the analysis state is particularly valuable when working with large volumes of data or developing complex longitudinal studies of scientific evolution.

3.5.3. CiteSpace: analysis of temporal patterns

CiteSpace stands out as a tool that specializes in analyzing temporal patterns and detecting emerging trends in the scientific literature. Developed by Dr. Chaomei Chen, this software allows users to map the evolution of scientific domains using advanced data mining techniques and dynamic visualization. Its ability to identify citation bursts and turning points in scientific development makes it an invaluable tool for technology foresight studies and diachronic research analysis.



Visible	Count	Central...	Year	Cited References
☑	38	0.00	2001	Schuster MA, 2001, NEW ...
☑	33	0.00	1997	Franz DR, 1997, JAMA-J A...
☑	31	0.00	2002	Galea S, 2002, NEW ENG...
☑	30	0.00	1999	Inglesby TV, 1999, JAMA-J ...
☑	29	0.00	1999	Henderson DA, 1999, JAM...
☑	27	0.00	1999	Henderson DA, 1999, SCI...
☑	26	0.00	1997	Torok TJ, 1997, JAMA-J AM...
☑	21	0.00	1999	North CS, 1999, JAMA-J A...
☑	17	0.00	1998	Hoffman B, 1998, INSIDE ...
☑	17	0.00	1997	Christopher GW, 1997, JA...
☑	15	0.00	1997	Simon JD, 1997, JAMA-J A...
☑	15	0.00	2002	Schlenger WE, 2002, JAM...
☑	15	0.00	1997	Tucker JB, 1997, JAMA-J A...
☑	13	0.00	1996	Mallonee S, 1996, JAMA-J ...
☑	12	0.00	1998	Okumura T, 1998, ACAD E...
☑	12	0.00	2000	Inglesby TV, 2000, JAMA-J ...
☑	12	0.00	2001	Wetter DC, 2001, AM J PU...
☑	12	0.00	2000	Macintyre AG, 2000, JAMA-...
☑	12	0.00	2001	Meltzer MI, 2001, EMERG I...
☑	12	0.00	1999	Alibek K, 1999, BIOHAZAR...
☑	12	0.00	1997	Zilinskas RA, 1997, JAMA-...
☑	12	0.00	1999	Brennan RJ, 1999, ANN E...
☑	11	0.00	1999	Keim M, 1999, ANN EMER...
☑	11	0.00	1997	Kaufmann AF, 1997, EME...
☑	10	0.00	1999	Pfefferbaum B, 1999, AM J...
☑	9	0.00	1999	*cdcp , 1999, MMWR-MOR...
☑	9	0.00	1997	Kolavic SA, 1997, JAMA-J ...
☑	9	0.00	1997	Holloway HC, 1997, JAMA-...
☑	9	0.00	2001	Waeckerle JF, 2001, ANN ...
☑	8	0.00	2001	Jernigan JA, 2001, EMER...
☑	8	0.00	2002	Inglesby TV, 2002, JAMA-J ...
☑	8	0.00	1999	Dixon TC, 1999, NEW EN...
☑	8	0.00	1998	Carter A, 1998, FOREIGN ...
☑	8	0.00	2001	Arnon SS, 2001, JAMA-J A...
☑	7	0.00	1996	Okumura T, 1996, ANN E...
☑	7	0.00	2001	Dennis DT, 2001, JAMA-J ...
☑	7	0.00	1999	Kortepeter MG, 1999, EME...
☑	7	0.00	2000	Khan AS, 2000, LANCET, ...
☑	7	0.00	1998	Okumura T, 1998, ACAD E...
☑	7	0.00	2002	Yehuda R, 2002, NEW EN...
☑	7	0.00	2002	Breman JG, 2002, NEW E...
☑	7	0.00	1999	Richards CF, 1999, ANN E...
☑	7	0.00	2002	Kaplan EH, 2002, P NATL ...
☑	6	0.00	1999	Harvey AG, 1999, J CONS...
☑	6	0.00	2000	Pfefferbaum B, 2000, PSY...
☑	6	0.00	1994	*am PSYCHASS, 1994, DI...
☑	6	0.00	2001	Inglesby TV, 2001, CLIN I...
☑	6	0.00	1998	Sharp TW, 1998, ANN EM...
☑	6	0.00	2000	*cdcp , 2000, MMWR-MOR...
☑	6	0.00	1999	Terr LC, 1999, AM J PSYC...
☑	6	0.00	2002	Rotz LD, 2002, EMERG IN...
☑	6	0.00	2001	Barbera J, 2001, JAMA-J A...
☑	6	0.00	1999	Otoole T, 1999, EMERG IN...
☑	6	0.00	2002	Silver RC, 2002, JAMA-J A...
☑	5	0.00	1999	Pfefferbaum B, 1999, J AM...
☑	5	0.00	1999	Smith DW, 1999, J OKLAH...

```
CiteSpace, v. 6.1.R3 (64-bit) Basic
August 16, 2022 at 5:53:45 PM SGT
WoS: C:\Users\aaron\.citespace\Examples\WoS\Terrorism1990-2003\data
Timespan: 1996-2003 (Slice Length=1)
Selection Criteria: g-index (k=25), LRF=3.0, L/N=10, LBY=5, e=1.0
Network: N=385, E=1250 (Density=0.0169)
Largest CC: 244 (63%)
Nodes Labeled: 1.0%
Pruning: None
Modularity Q=0.6544
Weighted Mean Silhouette S=0.8823
Harmonic Mean(Q, S)=0.7515
```

Control Panel

Colormap | Burstness | Search | Clusters | **Labels** | Layout | Views

Keyword | Term | Overlay Labels

By Degree ☐ Show Frequency

Threshold 10

Font Size 15

Node Size 10

Node Labels

By Cluster ☐ Show Frequency

Threshold 2

Font Size 5

Node Size 1

Link Labels

☐ Show Link Labels ☐ Show Link Strengths

Font Size 8

Source: CiteSpace official website (https://www.java.com/en/download).
Figure 3.9. Main view of CiteSpace, with project graphically represented

Installation requires the Java Runtime Environment (https://www.java.com/en/download) to be installed beforehand and is done using an executable JAR file available on the official website (https://citespace.podia.com/). The process includes configuring memory parameters to optimize performance with large volumes of data. CiteSpace organizes its interface into modular panels that sequentially manage data import, analytical processing, and result visualization, maintaining a structured workflow. However, it has a considerable learning curve for novice users.

Temporal analysis is based on the construction of citation networks segmented by user-defined periods. The software applies community-detection algorithms to identify thematic clusters and calculates centrality metrics to locate bridge publications across different areas of knowledge. The superimposition of consecutive networks allows the visualization of the emergence, convergence, or disappearance of lines of research over time, revealing the dynamics of scientific change.

The citation burst detection function identifies publications that experience sudden increases in citation frequency, signaling potentially revolutionary contributions or emerging research topics. These bursts are visualized by color-d rings on the network nodes, where the intensity and timing of the color indicate the magnitude and duration of the high citation period. This feature allows you to quickly identify works that have marked turning points in their field.

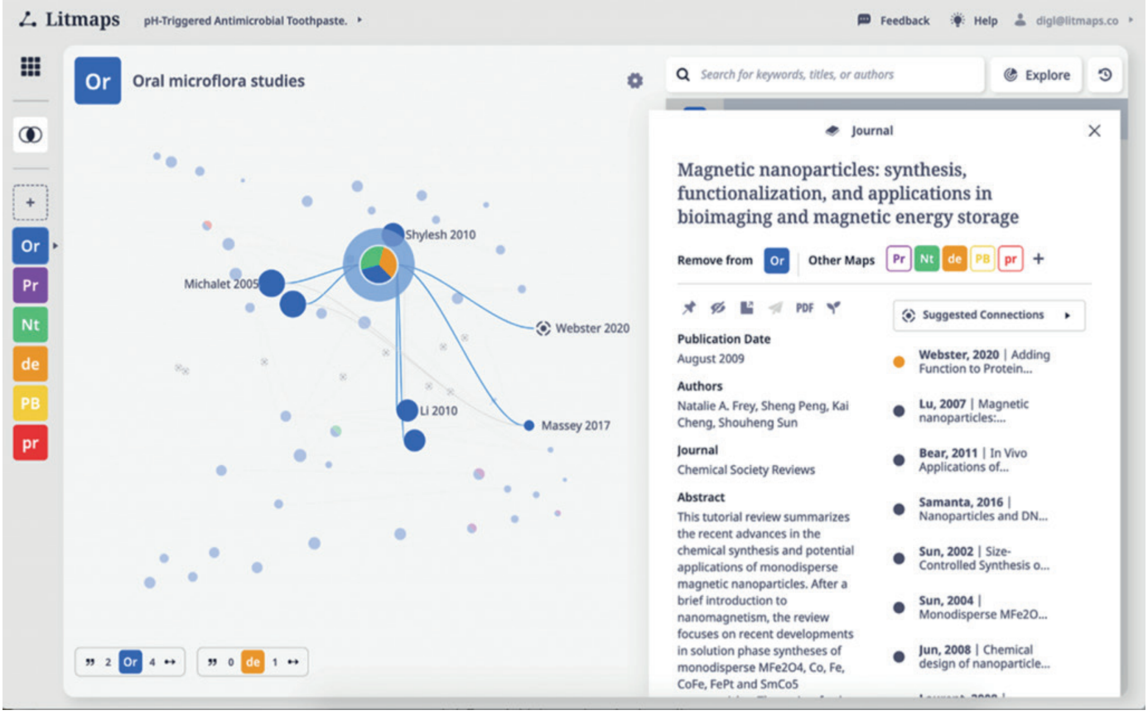
Scientific landscape visualizations use topographic models in which elevations represent connection density, and valleys indicate thematic discontinuities. CiteSpace generates evolutionary maps showing the drift of thematic clusters over time, complemented by trend lines that project future developments. These representations facilitate the identification of research opportunities and areas with growth and innovation potential.

The export of results includes detailed metric reports, animations of temporal evolution, and processed data for secondary analysis. CiteSpace generates tables of thematic clusters, including their fundamental publications, cohesion metrics, and representative labels, extracted using natural language processing algorithms. This comprehensive documentation supports the interpretation of detected patterns and facilitates the communication of complex findings on temporal dynamics in science.

3.5.4. Online tools (Litmaps, ResearchRabbit)

Online tools represent the latest evolution in bibliometric visualization, offering immediate accessibility and collaborative cloud-based functionalities. Litmaps and ResearchRabbit are emerging as innovative platforms that transform literature exploration through intuitive interfaces and advanced recommendation algorithms. These web solutions allow users to discover hidden connections between publications, track scientific developments in real time, and collaborate in geographically distributed research teams, marking a significant transition toward collaborative, real-time bibliometrics.

Litmaps (https://www.litmaps.com/) stands out for its ability to generate interactive citation maps that dynamically reveal relationships between publications. The platform allows users to visualize both the references cited in a seminal article and the subsequent citations of that article, creating knowledge networks that expand organically. Its "Seed Maps" algorithm automatically identifies the most relevant publications on a topic, while the alerts feature notifies users of new related research. This approach facilitates serendipitous literary exploration and the discovery of unexpected interdisciplinary connections.



Magnetic nanoparticles: synthesis, functionalization, and applications in bioimaging and magnetic energy storage

Publication Date
August 2009

Authors
Natalie A. Frey, Sheng Peng, Kai Cheng, Shouheng Sun

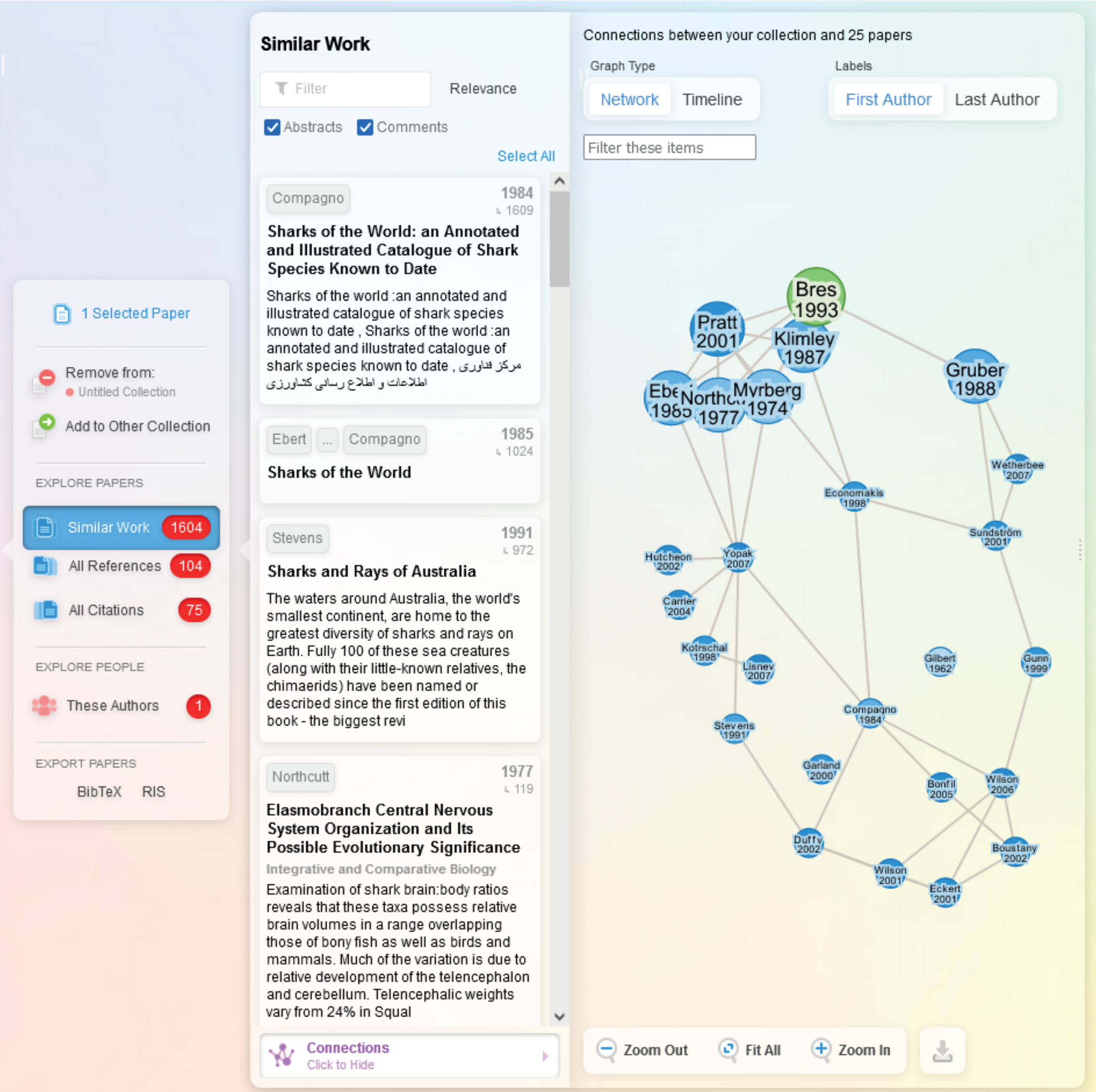
Journal
Chemical Society Reviews

Abstract
This tutorial review summarizes the recent advances in the chemical synthesis and potential applications of monodisperse magnetic nanoparticles. After a brief introduction to nanomagnetism, the review focuses on recent developments in solution phase syntheses of monodisperse MFe2O4, Co, Fe, CoFe, FePt and SmCo5

- Webster, 2020 | Adding Function to Protein...
- Lu, 2007 | Magnetic nanoparticles:...
- Bear, 2011 | In Vivo Applications of...
- Samanta, 2016 | Nanoparticles and DN...
- Sun, 2002 | Size-Controlled Synthesis o...
- Sun, 2004 | Monodisperse MFe2O...
- Jun, 2008 | Chemical design of nanoparticle...

Source: Litmaps official website (https://www.litmaps.com/).
Figure 3.10. Main view of Litmaps, with project graphically represented

ResearchRabbit (https://researchrabbitapp.com/) takes a different approach, functioning as "Spotify for academic research" through personalized recommendations based on reading preferences. The tool allows users to create collections of articles and receive increasingly refined suggestions through machine learning algorithms. Its timeline interface visualizes the historical evolution of research lines, while collaboration features enable users to share and comment on collections with research teams. This user-centered approach transforms literature exploration into a personalized and cumulative experience.



Compagno 1984 ↳ 1609

Sharks of the World: an Annotated and Illustrated Catalogue of Shark Species Known to Date

Sharks of the world :an annotated and illustrated catalogue of shark species known to date , Sharks of the world :an annotated and illustrated catalogue of shark species known to date , مرکز فناوری
اطلاعات و اطلاع رسانی کشاورزی

Ebert ... Compagno 1985 ↳ 1024

Sharks of the World

Stevens 1991 ↳ 972

Sharks and Rays of Australia

The waters around Australia, the world's smallest continent, are home to the greatest diversity of sharks and rays on Earth. Fully 100 of these sea creatures (along with their little-known relatives, the chimaerids) have been named or described since the first edition of this book - the biggest revi

Northcutt 1977 ↳ 119

Elasmobranch Central Nervous System Organization and Its Possible Evolutionary Significance

Integrative and Comparative Biology

Examination of shark brain:body ratios reveals that these taxa possess relative brain volumes in a range overlapping those of bony fish as well as birds and mammals. Much of the variation is due to relative development of the telencephalon and cerebellum. Telencephalic weights vary from 24% in Squal

Source: ResearchRabbit official website (https://researchrabbitapp.com/).

Figure 3.11. Main view of ResearchRabbit, with project graphically represented

Implementing these tools requires only a modern web browser and free registration, eliminating barriers related to installation and operating system compatibility. Both platforms synchronize data in the cloud, allowing access from multiple devices and automatic recovery of work sessions. Integration with reference managers such as Zotero and Mendeley facilitates the export of relevant discoveries, while connections to major databases ensure comprehensive coverage of up-to-date literature.

Practical limitations include dependence on a stable internet connection and considerations regarding research data privacy. The closed nature of their algorithms can make it challenging to validate the methodology behind the recommendations, and free versions typically impose limits on the volume of analysis. Nevertheless, their immediate usability and ability to reduce the time spent exploring the literature make them valuable additions to the modern bibliometric toolkit, particularly in the early stages of research and for researchers in training.

3.6. Selection of tools

The appropriate selection of bibliometric tools requires a systematic evaluation that considers the specific research objectives, available resources, and the technical characteristics of each software tool. This chapter provides a comprehensive comparative framework to guide the selection of tools across different analysis scenarios, highlighting both individual capabilities and potential synergies between complementary platforms. The optimal strategy often combines multiple tools in integrated workflows that leverage each tool's distinctive strengths.

For mapping scientific collaboration networks, VOSviewer is the preferred choice thanks to its layout algorithms optimized for visualizing relationships between authors and institutions. When the analysis requires identification of communities and structural patterns in large networks, the combination of Bibliometrix for statistical processing and VOSviewer for visualization produces particularly robust results. In projects that prioritize communicative clarity over analytical detail, online tools such as ResearchRabbit offer immediately understandable visualizations for non-specialist audiences.

CiteSpace is the most powerful tool for analyzing citation networks and temporal evolution, especially for detecting turning points and emerging trends in rapidly evolving scientific domains. When the main objective is to track the historical development of specific ideas, CitNetExplorer provides superior diachronic capabilities for visualizing knowledge trajectories. For technology foresight studies that require identifying areas of research opportunity, combining CiteSpace with Litmaps' burst analysis offers valuable complementary insights.

In contexts of individual or institutional scientific evaluation, Publish or Perish allows for quick calculations of fundamental indicators from Google Scholar. At the same time, Bibliometrix offers a more comprehensive analysis by integrating multiple data sources. For evaluation committees that require standardized, comparable metrics, the combination of both tools provides immediacy and analytical depth. In institutional environments with limited technical resources, online tools have the advantage of requiring minimal infrastructure and prior training.

Effective integration of bibliometric tools follows the principle of modularity, in which each tool contributes its specific strengths to a coherent workflow. A common strategy combines data extraction with Publish or Perish, statistical processing with Bibliometrix or PyBibx, and advanced visualization with VOSviewer or CiteSpace. This pipeline approach maximizes individual capabilities while mitigating the particular limitations of each tool, producing more comprehensive analyses than are possible with any single platform.

Technical interoperability is facilitated by standardized exchange formats such as RIS, BibTeX, and CSV, which enable data transfer between tools with minimal loss of information. Custom scripts in R or Python can automate these conversions, which is especially useful when integrating tools based on different technological ecosystems. Meticulous documentation of the parameters used at each stage ensures the reproducibility of the integrated analysis and facilitates the identification of discrepancies between results obtained with different tools.

Integration strategies must consider both technical aspects and available human resources. Teams with programming expertise can implement highly customized workflows that combine Bibliometrix with complementary analyses in R or Python. For groups with limited technical capabilities, integrating tools with graphical interfaces, such as VOSviewer and Publish or Perish, offers an effective balance between analytical capabilities and practical usability. Cross-training in complementary tools is a strategic investment that significantly expands the research team's analytical capabilities.

Recap

- Bibliometric tools enable the extraction, processing, visualization, and analysis of scientific data from specialized databases.
- Their systematic use improves the accuracy and reproducibility of bibliometric studies.
- There are three main types of tools:
  - Data retrieval and extraction.
  - Statistical analysis and processing.
  - Scientific information visualization and mapping.
- The most widely used bibliographic databases are Web of Science (WoS), Scopus, Dimensions, PubMed, Lens, and Google Scholar.
- Web of Science offers classic indicators such as the Journal Impact Factor and is essential for longitudinal studies.
- Scopus, from Elsevier, includes the CiteScore indicator and the SCImago Journal Rank (SJR) system.
- Google Scholar is a valuable open-source resource for broad coverage analysis, though with less quality control.
- Notable analysis tools include VOSviewer, Bibliometrix (R), CiteSpace, SciMAT, and Gephi.
- VOSviewer is widely used to construct maps of word co-occurrence, co-authorship, and co-citation.
- Bibliometrix, an R package, enables advanced statistical analysis and exports to Biblioshiny, its interactive web interface.
- CiteSpace focuses on detecting trends and thematic clusters through temporal citation networks.
- SciMAT is used for evolutionary analysis of scientific production and for visualizing the dynamics of topics over time.
- Gephi allows complex collaboration and co-citation networks to be represented with a high degree of visual customization.
- Statistical and text mining tools (Excel, SPSS, R, Python) are also used to process exported data.
- The combination of several tools enhances the validity of the analysis and enables complementary perspectives.
- The selection of the tool depends on the study's objective, sample size, and desired level of detail.
- It is essential to maintain transparency and traceability throughout the data

extraction and cleaning processes.

- Bibliometric visualization facilitates the interpretation of results and scientific communication.
- Tools must be used with consistent ethical and methodological criteria, avoiding manipulation of indicators.
- Mastery of these tools is an essential skill for researchers, analysts, and science managers.

Self-assessment questions

1. What are the three main categories of bibliometric tools?
2. What functions do databases such as Web of Science and Scopus perform?
3. What indicator mainly characterizes Web of Science?
4. What metrics system does Scopus offer in addition to CiteScore?
5. What are the advantages and limitations of Google Scholar for bibliometric analysis?
6. What is VOSviewer used for in scientific network studies?
7. What features differentiate Bibliometrix from other tools?
8. What does CiteSpace contribute to the study of thematic trends?
9. How important is visualization in bibliometric analysis?
10. Why is it essential to document data extraction and analysis processes transparently?

BIBLIOGRAPHY

1. Moed HF. Citation Analysis in Research Evaluation. Dordrecht: Springer; 2005. DOI: 10.1007/1-4020-3714-7

2. Thelwall M. Web Indicators for Research Evaluation: A Practical Guide. San Rafael (CA): Morgan & Claypool; 2016. DOI: 10.2200/S00733ED1V01Y201602ICR048

3. Sugimoto CR, Larivière V. Measuring Research: What Everyone Needs to Know. Oxford: Oxford University Press; 2018. ISBN: 9780190640125. https://global.oup.com/academic/product/measuring-research-9780190640125

4. Ding Y, Rousseau R, Wolfram D. Measuring Scholarly Impact: Methods and Practice. Cham: Springer; 2014. DOI: 10.1007/978-3-319-10377-8

5. Börner K, Chen C, Boyack KW. Visualizing Knowledge Domains. In: Cronin B, Sugimoto CR, editors. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. Cambridge (MA): MIT Press; 2014. p. 197–228.DOI: 10.7551/mitpress/9780262026792.003.0010

6. Chen C. CiteSpace: A Practical Guide for Mapping Scientific Literature. Hauppauge (NY): Nova Science Publishers; 2016. ISBN: 9781634842847

BIBLIOGRAPHIC REFERENCES

1. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics. 1 de noviembre de 2017;11(4):959-75.

2. Pereira V, Basilio MP, Santos CHT. PyBibX – a Python library for bibliometric and scientometric analysis powered with artificial intelligence tools. Data Technologies and Applications. 2025;59(2):302-37.

Chapter 4 / Capítulo 4

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch04

Data Collection / Obtención de Datos

From Chaos to Organized Knowledge

4.1. Search Strategies for Bibliometric Studies

In bibliometrics, the quality of results depends fundamentally on a rigorous, well-planned search strategy. A poor approach at this initial stage can compromise the entire study, generating significant biases, omitting essential literature, or incorporating irrelevant material that distorts the analysis. The consequences of inadequate design are dire in this field, where the integrity of the findings depends on the comprehensiveness and accuracy of the documentary corpus retrieved.

This chapter offers a systematic methodological framework for developing optimal search strings on the leading bibliographic platforms (Web of Science, Scopus, PubMed, among others). The approach presented combines theoretical principles with practical applications.

- Advanced techniques for formulating balanced search equations.
- Validation strategies to ensure complete coverage without sacrificing accuracy.
- Concrete examples drawn from various scientific disciplines.
- Frequent errors and how to avoid them, based on analysis of real cases.

The proposed method emphasizes the need to adapt each search to the particularities of the documentary system of each field of study, considering factors such as specific disciplinary terminology, the indexing characteristics of each database, and the biases inherent in different information retrieval systems.

Through this structured approach, researchers will be able to build robust document sets that serve as a reliable basis for advanced bibliometric analysis, from scientific mapping studies to impact assessments. The methodology presented has been empirically validated across multiple research projects, demonstrating its effectiveness in producing representative, replicable results across various areas of knowledge.

4.1.1. Designing the search strategy

Defining the scope: the basis of an effective bibliometric search

The first critical step in any bibliometric study is to establish the research's scope and objectives precisely. This delimitation process requires answering fundamental questions that will determine the entire subsequent methodological design.

The central research question must be clearly formulated, as it will define the type of analysis to be performed. Is the objective to identify temporal trends in a field? To evaluate the impact of certain institutions? To map international collaboration networks? Each of these purposes will require different search strategies.

Inclusion criteria are essential filters for ensuring the relevance and manageability of the results:

- The time period must be justified according to developments in the field (e.g., in emerging areas, the last 5 years may be sufficient; in established disciplines, a decade or more may be necessary).
- The types of documents selected (articles, reviews, patents, theses) must be aligned with the communication practices of each discipline.
- The decision on languages involves a balance between comprehensiveness (including

contributions in multiple languages) and feasibility (the predominance of English in many scientific areas).

Applied example: To study *"Scientific production on artificial intelligence applied to medicine between 2015-2025,"* an appropriate delimitation could specify:

- Only full articles in peer-reviewed journals
- Documents indexed in major databases (WoS, Scopus)
- Publications in English (given its predominance in these technical disciplines)
- Exclusion of patents and gray literature to maintain focus on basic research

This planning phase, although often underestimated, is crucial. A large part of the errors in bibliometrics originates from an imprecise definition of the initial scope. A clear delimitation subsequently avoids the inclusion of documentary noise or the omission of key literature, problems that can compromise the validity of the findings.

Select databases

Given the planned scope, the database to be used for the search must be selected, which requires basic knowledge of each database's content and strengths. Each database has advantages and biases:

Table 4.1. Coverage, filters, and limitations of the main scientific databases

Database	Coverage	Useful filters	Limitations
Scopus	+25,000 journals, strong in STEM	Integrated citation analysis	Expensive, weak in humanities
Web of Science	+20,000 journals, historical core	Built-in h-index	Bias toward the US/Europe
PubMed	+30 million records in biomedicine	Filters by study type (clinical trials)	No complete citation data
Google Scholar	Free, includes books and gray literature	Semantic search	No analysis tools
Dimensions	Integrates publications and patents	Funding data	Uneven coverage by region

Database biases can be offset by using multiple databases (e.g., Scopus + PubMed).

4.1.2. Building search strings

Once the topic of the search has been clearly defined, the words most representative of the research should be selected.

Keywords and Boolean operators

The effectiveness of a bibliometric search depends on how key terms are combined using logical operators. The AND operator restricts results by connecting concepts that must appear simultaneously (example: *"machine learning"* AND *"healthcare"* retrieves only documents that mention both topics). Conversely, OR broadens the scope by including synonyms or conceptual variants (e.g., "*deep learning*" OR *"neural networks"*). NOT is used to exclude irrelevant topics, although it should be used with caution so as not to eliminate useful literature (e.g., "*cancer*" NOT *"breast cancer"* could omit critical cross-sectional studies).

Quotation marks ensure exact searches for complete phrases, avoiding scattered results ("climate change" is more accurate than climate AND change). Wildcards () capture morphological variants of a common root (e.g., "education," "educative," "educational").

Example of an optimized string

("artificial intelligence" OR "AI") AND ("medical diagnosis" OR "clinical decision") NOT "robotics."

This query retrieves documents on AI applied to medical diagnosis, excluding those focused on robotics.

The above operators are common to most databases and search engines, but some databases add new ones to increase search specificity. In any case, the logic behind these operators is the same regardless of where they are implemented.

Advanced filters for precision

Often, the above is not enough to retrieve the required or targeted content, or it covers much more than planned. Databases allow you to refine searches using specific filters:

- By field: in Scopus, TITLE-ABS-KEY() searches the title, abstract, and keywords; in PubMed, TI/AB covers the title and abstract. Limiting the search to the title only (TITLE()) underestimates the relevant literature, as many key concepts may only appear in the abstract.
- By impact: filters such as CITES > 50 (WoS) identify highly cited works, useful for studies of scientific influence.
- By institution: AFFIL("Harvard University") locates the output of a specific university, essential for collaboration analysis.

Standard error and solution

Problem: using only TITLE() can cause relevant documents to be missed

Solution: always combine title, abstract, and keywords (TITLE-ABS-KEY() in Scopus or TI/AB/KW in WoS).

4.1.3. Specific techniques by database

In Scopus, advanced syntax allows precise filtering using operators such as TITLE-ABS-KEY(), which covers titles, abstracts, and keywords simultaneously. For example, a query such as TITLE-ABS-KEY("blockchain" AND "supply chain") AND PUBYEAR > 2015 retrieves only recent publications on blockchain applications in supply chains.

Web of Science offers unique features for institutional network analysis. Its TS=() syntax for search terms combined with AD=() for affiliations allows you to map collaborations between specific research centers. A query such as TS=("quantum computing") AND AD=("MIT" OR "Stanford") identifies the joint output of these universities in quantum computing.

For biomedical research, PubMed stands out for its controlled vocabulary system, MeSH (Medical Subject Headings). This hierarchical thesaurus resolves problems of synonymy and terminological ambiguity. A search such as "COVID-19"[Mesh] AND "Vaccines"[Mesh] AND "clinical trials"[Publication Type] ensures complete coverage of COVID-19 vaccine clinical trials, avoiding the biases of free-text searches. PubMed's precision makes it indispensable for systematic reviews in health sciences.

Each platform requires specific strategies to optimize precision and recall. Scopus and WoS prioritize control over terminology and integrated analytical capabilities, while PubMed offers standardized vocabulary, and Google Scholar provides documentary breadth at the expense of methodological transparency. The choice between them should be based on the objectives of the study: citation analysis and collaboration (WoS/Scopus), rigorous clinical reviews (PubMed), or identification of unconventional literature (Google Scholar). A robust research design often combines multiple databases to compensate for their respective limitations.

Recently, tools such as Dimensions have emerged as promising alternatives, integrating publications, patents, and datasets into a single environment with an open API. Their flexible syntax (e.g., search publications for "CRISPR" where year > 2020) and multidisciplinary coverage are redefining the standards for complex bibliometric searches. However, the widespread adoption of these new systems still faces interoperability challenges with traditional metrics.

In summary, the specific operators and filters for each database are:

Scopus
Boolean and advanced search operators

- AND, OR, AND NOT → Combine terms.
- → Truncation (e.g., comput searches for "computer," "computing," etc.).
- " " → Exact search (e.g., "machine learning").
- W/n → Words within n terms of distance (e.g., AI W/3 healthcare).
- PRE/n → First word before the second, with n terms away (e.g., digital PRE/2 transformation).

Advanced filters

- Year: PUBYEAR > 2020
- Document type: DOCTYPE(ar) (articles) or DOCTYPE(re) (reviews).
- Specific field:
    - TITLE-ABS-KEY("blockchain") → Search in title, abstract, or keywords.
    - AFFIL("Harvard University") → By affiliation.
- Citation indexes: REFERENCES() (e.g., REFERENCES(123456789)).

Combined example: TITLE-ABS-KEY("quantum computing") AND PUBYEAR > 2018 AND AFFIL("MIT") AND DOCTYPE(ar)

Web of Science (WoS)
Operators and syntax

- AND, OR, NOT → Standard operators.
- → Truncation (e.g., bio for "biology," "biotechnology").
- " " → Exact phrase.
- NEAR/x → Nearby terms (e.g., climate NEAR/2 change).
- SAME → Terms in the same field (useful for addresses: UNIV COLORADO SAME BOULDER).

Advanced filters

- WC categories: WC=("Computer Science").
- Year: YR=2020-2023.
- Document type: DT=("Article").
- Search by field:
    - TS=("neural network") → Subject (title, abstract, keywords).

  - AU=("Smith J") → Author.
  - OG=("Stanford University") → Organization.

Combined example: TS=("CRISPR" AND "gene editing") AND PY=(2020-2023) AND WC=("Biotechnology") AND DT=("Review")

PubMed

Operators and syntax

- AND, OR, NOT → Boolean operators.
- → Truncation (e.g., immun for "immune," "immunity").
- " " → Exact phrase.
- [Field] → Search by field (e.g., "COVID-19"[Title]).

Advanced filters (using tags)

- Article type: "review"[Publication Type].
- Date: "2020/01/01"[Date - Publication] : "2023"[Date - Publication].
- Specific fields:
  - "deep learning"[Title/Abstract].
  - "NIH"[Affiliation].
  - "1AU"[Author] → Lead author.
- Predefined filters:
  - "free full text"[Filter].
  - "clinical trial"[Filter].

Combined example: "artificial intelligence"[Title/Abstract] AND "diagnosis"[Title]) AND "2020"[Date - Publication] AND "systematic review"[Publication Type]

4.1.4. Validation and documentation: ensuring rigor in bibliometric searches

The validation phase is essential to ensure that the results of a bibliometric search are comprehensive, accurate, and reproducible. A rigorous process begins with a **pilot test**, where the first 50-100 records obtained are critically analyzed. This initial sampling allows frequent problems to be identified: overly broad terms that generate documentary noise, or, conversely, excessively restrictive search strings that omit relevant literature. Iteration is key here; each adjustment to the Boolean operators or term selection must be tested until an optimal balance between precision and sensitivity is achieved.

To ensure methodological transparency, the **PRISMA** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) **protocol** provides a standardized framework applicable to bibliometric studies. Its systematic approach requires detailed documentation of the entire selection process: the total number of records identified in each database, the exclusion criteria applied (duplicates, documents outside the temporal or thematic scope), and the decisions made at each stage of filtering. The PRISMA flowchart thus becomes a powerful visual tool that summarizes the process of refining the document corpus, allowing other researchers to evaluate the methodological soundness of the study.(1)

Detailed documentation should also include: the exact syntax of all search strings used (including discarded variants), the dates on which the queries were run (important given the dynamic nature of databases), and any known limitations in the coverage of the sources used. Clarity on all of these points is crucial from the initial phase of the research, as, as will be seen in subsequent chapters, all of this must be detailed in the final report.

Experienced researchers often supplement this process with **external validation**, consulting subject matter experts to verify that the search has not omitted seminal contributions, or comparing their results with existing systematic reviews in the field. This additional step helps to detect possible biases in the search strategy that could affect the study's conclusions. The combination of internal validation (using standardized protocols) and external validation (through expert judgment) is the best guarantee of producing robust and reliable bibliometric results.

This methodical approach transforms the bibliographic search from a mere technical procedure into a rigorous research process, where each decision is justified and documented. The time invested in this critical phase directly affects the quality and credibility of subsequent findings, helping avoid the biases and limitations that often affect bibliometric studies conducted without this systematic approach.

This progressive reduction must be meticulously documented, specifying the reasons for exclusion at each stage. Complete traceability of the process not only strengthens the internal validity of the study but also facilitates its replication or future updating.

4.1.5. Common errors and solutions

An overly broad search, yielding more than 10,000 records, is often unmanageable for detailed bibliometric analysis and may include substantial documentary noise. This problem frequently occurs when using generic terms without adequate filters. The solution is to incorporate strategic restrictions such as time limits (e.g., last 10 years), filtering by document type (research articles only), or focusing on specific fields (title and abstract rather than full text).

At the other extreme, an overly narrow search can miss seminal contributions by using overly specific terminology. This is particularly problematic in emerging fields where concepts have not yet standardized their nomenclature. To avoid this, it is essential to build search strings that incorporate synonyms and terminological variants using OR operators, in addition to consulting disciplinary thesauri such as MeSH in PubMed or Scopus controlled terms.

A common mistake in specific disciplines is to ignore books and conference proceedings, which creates a systematic bias against areas such as the humanities and computer science. Traditional databases (WoS/Scopus) have limited coverage in these formats, so the search should be supplemented with Google Scholar for books or Dimensions for proceedings. In interdisciplinary studies, this omission can lead to missing much of the relevant literature.

Failure to document the search strategy compromises the study's reproducibility, a fundamental principle of research. The solution involves systematically saving not only the final chain but all its iterations, specifying the platform used, the search date, the filters applied, and the number of results obtained at each step. Standard formats such as PRISMA diagrams or systematic review protocol templates can be adapted for this purpose. Tools such as Zotero or Mendeley allow search records to be exported with complete metadata for later auditing.

These problems and their solutions illustrate how the design of bibliometric searches is an iterative process that requires a balance between comprehensiveness and precision. Experience shows that dedicating 30 % of the total study time to refining and validating the search strategy can significantly improve the quality of the final results. Incorporating expert peer reviews in the methodological design phase helps identify potential omissions or biases before executing the final searches.

The growing use of machine learning techniques for automated screening of large volumes of results is transforming this process, but the initial construction of accurate search strings remains essential. Platforms such as ASReview or Rayyan can then assist in the selection phase, but they depend critically on a well-designed document retrieval strategy from the outset. This synergy between traditional methods and new technologies paves the way for more efficient and comprehensive bibliometric searches.

4.2. Alternative tools for bibliometric search: broadening horizons

Although Scopus and Web of Science remain benchmarks in bibliometric analysis, there are valuable alternatives that democratize access to scientific information. Google Scholar, with its extensive coverage of gray literature, along with tools such as Publish or Perish and advanced reference managers, offers unique possibilities for researchers with limited resources or who work with unconventional formats.

4.2.1. Google Scholar: practical advantages and limitations

The main strength of Google Scholar lies in its ability to index content that is often left out of commercial databases: from preprints on arXiv to doctoral theses and working papers. This comprehensiveness makes it particularly useful for studies in the humanities or emerging disciplines where formal publication can be slower. However, its systematic use presents significant technical challenges. The approximate limit of 100 queries per day before IP blocking forces the use of careful sampling strategies, while the lack of structured metadata complicates advanced bibliometric analysis.

For researchers with technical skills, Python’s scholarly library offers a balance between automation and compliance with terms of service. The basic process involves: installation via pip, configuration of specific queries, and controlled data extraction. A script might include:

~~~~
```python
import csv
from scholarly import scholarly
search_query = scholarly.search_pubs(“blockchain in healthcare”)
with open(‘results.csv’, ‘w’, newline=’’, encoding=’utf-8’) as file:
    writer = csv.writer(file)
    writer.writerow([“Title”, “Authors”, “Year”, “Citations”])
    for i in range(50):   Responsible limit
        try:
            pub = next(search_query)
            writer.writerow([
                pub[‹bib›].get(‹title’, ‘’),
                pub[‹bib›].get(‹author’, ‘’),
                pub[‹bib›].get(‹year’, ‘’),
                pub.get(‘num_citations’, 0)
            ])
        Except StopIteration:
            break
```
~~~~

This approach includes key safeguards: record limits, error handling, and CSV storage for later analysis. For non-programmers, extensions such as Scholarcy offer basic extraction functionality directly from the browser, albeit with less customization.

Academic data scraping must follow clear principles: intervals between queries that simulate human behavior, respect for each site's robots.txt, and exclusive use for research purposes. It is crucial to remember that many documents on Google Scholar are protected by copyright, so mass extraction of full texts without permission constitutes an ethical violation. Best practices recommend:

- Limiting extractions to basic metadata (title, authors, citations).
- Storing only the information necessary for analysis.
- Properly citing all retrieved sources.
- Periodically checking the platforms' terms of service.

These alternative tools work best as complements, not replacements, to traditional databases. A robust workflow could begin with systematic searches in Scopus/WoS, followed by a sweep of Google Scholar to capture additional literature, and end with tools such as Zotero or Mendeley for organization and deduplication. This methodological triangulation mitigates the inherent biases of each platform while maximizing document coverage.

4.2.2. Harzing's Publish or Perish

Publish or Perish (PoP) not only extracts data from Google Scholar but also integrates with key academic databases such as Scopus, Web of Science (WoS), and PubMed, significantly expanding its usefulness for researchers who require more structured analysis. This multi-platform capability makes it a versatile solution for comparative bibliometric studies, allowing impact metrics to be compared across different sources and detecting possible discrepancies in citation coverage. Chapter 3 discussed in detail how to install this tool and its advantages and features. In general, to perform a search with it, follow these steps:

1. Download and installation:
   - Available at https://harzing.com/resources/publish-or-perish (Windows/Mac/Linux).

2. First search:
   - Open the software and click on "New Query."
   - Enter search terms (e.g., "artificial intelligence education").

3. Export results:
   - Go to File > Save As and choose CSV/Excel for analysis in Bibliometrix or Excel.

4.2.3. Reference managers for advanced bibliometric analysis

Reference managers such as Zotero and Mendeley have evolved beyond simple bibliographic organizers into powerful tools for extracting and mass-processing bibliometric data. Their ability to capture structured metadata directly from academic portals or PDF files makes them indispensable in quantitative research workflows.

Zotero stands out for its seamless integration with databases such as Scopus and Web of Science. Using its browser connector, researchers can import hundreds of references with a single click, preserving critical information such as DOIs, institutional affiliations, and citation fields. The CSV export process (via the context menu) generates tables ready for analysis in specialized software. However, care is required when selecting custom fields to ensure that all relevant metadata is included.

Mendeley offers distinct advantages, particularly in handling existing PDF collections. Its document recognition engine can extract metadata even from files without embedded information, although with variable accuracy.

The "Verify Document Details" feature allows you to correct errors commonly found in older articles or less established journals. For bibliometric projects, its ability to export in RIS format is especially valuable, as this standard preserves relationships between documents that tabular formats such as CSV might lose.

A recurring challenge with both tools is the inconsistent quality of metadata. Recent studies show that up to 15 % of articles in reference managers require manual correction of basic fields, such as the year of publication or the complete list of authors. Here, APIs such as CrossRef become essential add-ons.

The Python example shown (using the habanero library) illustrates how to automate the retrieval of missing metadata via queries based on titles or text fragments. This hybrid approach—combining automatic extraction with programmatic verification—can significantly improve the efficiency of preparing bibliometric data.

Advanced technical considerations arise when working with large volumes of references. Zotero allows custom scripts to be run through its internal API (Zotero API), enabling tasks such as:

- Automatic normalization of institutional names.
- Detection and merging of duplicates based on semantic similarity.
- Extraction of citations between documents through reference analysis.

Meanwhile, Mendeley Desktop (unlike its web version) provides direct access to its local SQLite database, enabling complex queries beyond the limitations of the graphical interface. Researchers with SQL knowledge can thus perform advanced data transformations before final export.

The choice between these tools often depends on the specific workflow. Zotero shines in collaborative projects thanks to its cloud synchronization and granular permissions, while Mendeley may be preferred in environments where managing large volumes of PDFs is a priority. Both, however, share the potential to serve as a bridge between initial literature collection and sophisticated bibliometric analysis on platforms such as VOSviewer or Bibliometrix.
These managers can be used to extract data from PDFs or even from the web for later use in bibliometric studies.

4.3. Export formats in bibliometrics: interoperability and analysis

Selecting the export format in bibliometrics is a critical decision that directly affects the quality and depth of subsequent analysis. The RIS (Research Information Systems) format, initially developed for academic databases, stands out for its ability to preserve complex metadata such as citation relationships and hierarchical authorship structures. This format maintains a unique balance between human readability and automated processing, making it particularly valuable when working with network analysis tools such as VOSviewer. Its standardized tag structure (e.g., AU for authors or PY for year of publication) enables a smooth transition between platforms without loss of critical information.

BibTeX, on the other hand, offers distinct advantages in environments where technical academic writing is paramount. Its minimalist syntax and native integration with LaTeX systems make it irreplaceable for researchers who need to manage references while writing. However, this same simplicity becomes a limitation when addressing advanced bibliometric analysis, where richer metadata on citations, institutional affiliations, or impact indicators is required. Modern tools such as PyBibX have expanded their usefulness by enabling some bibliometric processing, but they still need additional transformations to achieve the level of detail provided by other formats.

The CSV format represents the most direct bridge to advanced quantitative analysis. Its tabular structure is ideally suited to statistical environments such as R (via Bibliometrix) or Python (with Pandas), where manipulating large volumes of data requires flexibility. However, this same adaptability carries risks: automatic exports from some platforms can flatten complex data structures, such as author lists or institution hierarchies, into text strings that require further processing. The key to leveraging CSV in bibliometrics is to ensure that the initial export captures all relevant metadata in well-defined columns.

4.3.1. Conversion between formats: tools and uses in bibliometric software

The reality of contemporary bibliometric work often requires navigating between multiple formats, each optimized for different stages of the research flow. Conversions between RIS, BibTeX, and CSV are not mere technical translations, but processes that can expand or limit analytical possibilities. Tools such as Zotero perform these transformations acceptably for basic uses, but large-scale projects require more sophisticated approaches.

In today's bibliometric software ecosystem, each primary tool has developed specific preferences and capabilities for handling these formats. VOSviewer, for example, has optimized its engine to process RIS files while preserving document relationships, which is essential for building accurate co-citation networks. CiteSpace, with its focus on temporal analysis, demands data structures that preserve citation chronologies, which plain-text formats derived from the Web of Science naturally provide.

Bibliometrix in R sits at the most flexible end of this spectrum, capable of ingesting multiple formats but with a clear advantage when fed RIS. Its `convert2df()` function not only interprets standard metadata, but also applies cleaning and normalization algorithms that prepare the data for complex analyses such as co-word mapping or the calculation of international collaboration indicators.

For researchers working with custom pipelines, Python offers libraries such as PyBibX that bridge the limitations of BibTeX for bibliometric analysis. These tools, for example, allow the extraction of semantic networks from sets of references intended initially for the composition of academic documents. However, the process often requires intermediate transformations, especially when the data comes from multiple sources with different metadata standards.

Decisions about formats and conversions directly impact the validity of bibliometric findings. An illustrative case is the analysis of institutional collaborations. While RIS typically preserves the complete structure of affiliations, automatic CSV conversions can fragment this information into separate columns, introducing artifacts in network maps generated by tools such as CitNetExplorer.

Experience recommends adopting a layered approach to format management: in the initial collection phase, RIS is usually the safest option, ensuring that no relevant metadata is lost during database exports. For the cleaning and normalization stages, CSV provides the flexibility needed to manipulate large volumes of data with statistical tools. Finally, in the visualization and specialized analysis phase, many researchers choose to return to formats such as RIS or to structures native to the bibliometric software used.

This iterative cycle underscores a fundamental principle: in advanced bibliometrics, data exchange formats are not mere passive containers, but active components that condition analytical possibilities. The growing interoperability between tools, driven by standards such as OpenAPI and formats such as JSON-LD, promises to simplify these workflows in the future. However, until these innovations are widely adopted, a thorough knowledge of RIS, BibTeX, and CSV remains an essential skill for any researcher seeking to extract meaningful insights from bibliographic data.

All the information is already there, but now what to do with it? The following section will address the entire data processing workflow to obtain high-level results and interpretations.

Recap

- Obtaining bibliometric data is the first practical phase of a study and defines the validity of all subsequent analysis.
- The primary sources of information are Scopus, Web of Science (WoS), and PubMed, due to their coverage, structure, and metadata exportability.
- In Scopus, the search is optimized using Boolean operators (AND, OR, NOT), and results are filtered by year, country, area, or affiliation, as well as by SOURCE-ID codes.
- In Web of Science, the Core Collection, Emerging Sources, and subject categories are used to refine results.
- In PubMed, controlled MeSH terms, clinical limits, and combination strategies between title/abstract/keywords are used.
- Alternative tools expand data collection when traditional databases are insufficient or restricted. These include:
  - Scholar (ethical scraping) → helpful in expanding coverage, but caution is advised due to noise and duplicates.
  - Harzing's Publish or Perish → allows you to extract and analyze Google Scholar metrics with customizable Google filters.
  - Reference managers (Zotero, Mendeley) → facilitate bulk export of metadata and format conversion.
- The most commonly used formats for exporting records are:
  - RIS → compatible with most bibliographic software.
  - BibTeX → standard for analysis in R or Python.
  - CSV → allows direct manipulation in Excel or statistical software.
- It is essential to maintain format consistency throughout the process, avoiding loss of fields (authors, affiliation, DOI, country, year).
- It is recommended to retain the original version of each dataset and document the data acquisition process (date, source, filters, and the final number of records).
- Ethical data collection prohibits practices such as unauthorized mass scraping or the use of restricted data.
- Reproducibility is promoted by clearly describing the search criteria and publishing the dataset or code used.
- The chapter links data collection with the following methodological stages: cleaning, analysis, and communication of results.

- In this phase, the relevant data are defined: bibliographic variables (author, year, journal, country, citations, affiliation, document type).
- The most common errors at this stage are: duplicates, incomplete records, errors in author names, and confusion between sources.
- Quality control includes verifying consistency, eliminating duplicates, and standardizing names through normalization.
- Complete documentation of the search process increases the transparency and credibility of the bibliometric study.
- It is recommended to accompany this phase with screenshots or reproducible scripts (e.g., in Google Colab or RMarkdown).
- The chapter emphasizes that without a solid database, subsequent metrics and visualizations lose reliability.
- In summary, this part teaches how to build the “empirical heart” of the study: an ethical, clean, complete, and standardized bibliographic database.

Self-assessment questions

1. Why is data collection considered the critical phase in a bibliometric study?
2. What characteristics differentiate Scopus, Web of Science, and PubMed as data sources?
3. What role do Boolean operators play in bibliographic searches?
4. What are the advantages and limitations of using Google Scholar through ethical scraping?
5. What is the purpose of Harzing’s Publish or Perish program in bibliometrics?
6. What is the difference between RIS, BibTeX, and CSV formats when exporting records?
7. What best practices should be followed to ensure the reproducibility of bibliographic searches?
8. What are the main errors that can affect the quality of the bibliometric dataset?
9. What ethical criteria should be respected when collecting scientific data?
10. Why are documentation and standardization of the dataset essential before bibliometric analysis?

BIBLIOGRAPHY

1. Moed HF. Citation analysis in research evaluation. Dordrecht: Springer; 2005. doi: 10.1007/1-4020-3714-7

2. Thelwall M. Web indicators for research evaluation: A practical guide. San Rafael (CA): Morgan & Claypool Publishers; 2016. doi: 10.2200/S00733ED1V01Y201602ICR048

3. Sugimoto CR, Larivière V. Measuring research: What everyone needs to know. Oxford: Oxford University Press; 2018. https://global.oup.com/academic/product/measuring-research-9780190640125

4. Ding Y, Rousseau R, Wolfram D, editors. Measuring scholarly impact: Methods and practice. Cham (Switzerland): Springer; 2014. doi: 10.1007/978-3-319-10377-8

5. Cronin B, Sugimoto CR, editors. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. Cambridge (MA): MIT Press; 2014. doi: 10.7551/mitpress/9780262026792.001.0001

6. Cooper HM, Hedges LV. Research synthesis and meta-analysis: A step-by-step approach. 5th ed. Los Angeles (CA): SAGE Publications; 2019.

BIBLIOGRAPHIC REFERENCES

1. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews. 2022; 18(2):e1230.

Part III / Parte III

ADVANCED ANALYSIS

ANÁLISIS AVANZADO

AG EDITOR

Chapter 5 / Capítulo 5

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch05

Bibliometric Indices / Índices Bibliométricos

5.1. Classification

Bibliometric indices are at the heart of quantitative analysis in science, technology, and innovation studies, providing standardized measures for evaluating the impact, productivity, and influence of scientific output. Their correct classification and interpretation are essential for drawing valid conclusions and avoiding the frequent errors that arise from mechanical applications without adequate contextualization. This chapter presents a comprehensive framework for understanding the taxonomy, calculation, and interpretation of the leading bibliometric indicators used in contemporary scientific evaluation.

Scientific Production Indicators quantify the volume of research results generated, providing an initial basis for evaluating scientific activity. The most basic metric is the total number of publications. A low count may be a symptom of limited productivity or a deliberate strategy to publish exclusively in high-impact journals, which involve more extensive review processes. Conversely, a high volume suggests high productivity, but it can also signal a possible fragmentation of results, where the findings of a single study are divided across multiple articles. The standards for what is considered “high” or “low” vary widely across disciplines and career stages.

To refine the measurement, the individual productivity index adjusts a researcher’s contribution score by accounting for factors such as their position on the author list and the type of contribution. A low value on this index indicates majority participation as a co-author in secondary positions, suggesting a supporting role in projects. A high value, on the other hand, denotes frequent primary authorship (as first or last author), reflecting sustained intellectual leadership. Although there is no universal threshold, in many disciplines, a value above 0,7 on normalized scales is considered a substantial contribution and a sign of leadership.(1)

The collaboration index measures teamwork by averaging the number of co-authors per publication. A low value, close to 1, is characteristic of traditionally solitary fields such as philosophy or the humanities. A high value, which can exceed ten in areas such as particle physics or genomics, reflects eminently collaborative research. The interpretation of this indicator must take disciplinary norms into account, as multiple authorship responds to scientific traditions and methodological needs that differ significantly between fields of knowledge.(2)

Citation-based impact indicators assess the relative influence and importance of scientific publications by measuring how they are received and used by the academic community through the references they receive.

The total number of citations provides a crude measure of cumulative impact. However, its interpretation requires temporal and disciplinary normalization. A low value may indicate research that is uninfluential or highly specialized, while a high value suggests fundamental contributions that have resonated in the field. There is no universal standard, as several citations considered low in immunology could be exceptionally high in philosophy.

The h-index is a bibliometric indicator that balances a researcher’s productivity with the impact of their work. It is calculated by ranking an author’s publications by the number of citations received, in descending order.

The h-index corresponds to the point at which the order number of an article (h) coincides with the number of citations that article has received. For example, an h-index of 10 means that the author has 10 articles that have been cited at least 10 times each.

A low value may indicate a fledgling scientific career or a specialization in an area of study with a generally low citation rate. Conversely, a high h-index usually reflects an established career and the consistent production of work that has had a significant impact in its field.

To make a fairer comparison between researchers in different fields, the discipline-corrected h-index is used. It is calculated in several steps. First, the researcher's primary discipline is identified. Then, the average h-index for all researchers in that discipline is obtained from a standardized bibliographic database.

Finally, the researcher's individual h-index is divided by this reference average. The result is a value that indicates whether their impact is above or below the average for their specialized field.

As a reference, Hirsch proposed that, for active scientists, an h-index approximately equal to the number of years in their career (m) is characteristic of a "successful" researcher. A value around 2m would correspond to "outstanding researchers," while an h ≈ 3m would be associated with "truly unique scientists," always considering the particularities of each discipline.(3)

To complement this, the g-index gives greater weight to exceptionally cited publications, being more sensitive to the presence of seminal works within an author's output. A g-index value that is significantly higher than the h-index suggests that the researcher has one or more highly cited articles with an extraordinary impact. Finally, the m-index adjusts the h-index for years of research career, providing a measure of annualized impact productivity. An m value greater than 1 is considered very good, as it indicates that the researcher generates, on average, more than one "h" article per year.(3)

The m index adjusts the h index for years of research experience, with lower values indicating decreasing impact and higher values indicating sustained productivity.

Scientific Collaboration Indicators analyze the networks and patterns of cooperation between researchers, institutions, and countries, revealing the social dynamics behind knowledge production. The cooperation index calculates the percentage of publications with multiple authors. Low values may indicate individualistic work or be characteristic of traditionally solitary fields. Conversely, high values, often above 80 % in the experimental sciences, reflect eminently collaborative research, typical of modern science where collective authorship is the norm for addressing complex problems.

The international collaboration index measures the proportion of works with foreign co-authors. A low value is often indicative of a certain scientific isolation or a research approach with mainly local relevance. A high value, on the other hand, demonstrates active integration into global knowledge networks and is often correlated with greater impact. In institutional evaluations, the aim is usually to have this index exceed 30-40 %, with values above 50 % considered high. Similarly, the institutional collaboration index assesses the diversity of affiliations in publications, indicating the ability to establish strategic alliances beyond the institution.

Source Quality Indicators focus on evaluating the prestige and influence of scientific dissemination channels, such as academic journals, assuming that the quality of the publication vehicle reflects, to a certain extent, the quality of the articles it contains. The Impact Factor (IF), calculated annually in the Journal Citation Reports, measures the average frequency with which articles in a journal are cited in a specific period. Low values may correspond to highly specialized or regional journals. High values, on the other hand, typically reflect leading journals in their fields.

To overcome the limitations of IF, alternative metrics have been developed. The SCImago Journal Rank (SJR) reflects the prestige of citing journals and is more sensitive to the quality of citations received. CiteScore uses a broader citation window. Source Normalized Impact per Paper (SNIP) corrects for differences in citation practices between disciplines, where values below 1 indicate an impact below the average for their field and values above 1 indicate an effect above the average. These indicators should be used critically, avoiding inappropriate extrapolations at the article or individual researcher level.[4]

Influence and Leadership Indicators assess the specific role and actual contribution of a researcher within scientific networks, going beyond metrics of volume or raw impact. The individual h-index excludes self-citations to strictly measure external impact and recognition by peers outside the researcher's immediate circle. A low value on this index may indicate a high dependence on self-citations to maintain a citation profile or a limited impact. A high value, on the other hand, reflects strong external recognition.

The leadership index calculates the proportion of publications in which the researcher appears as a corresponding or first author, roles that typically denote a primary intellectual contribution. Low values suggest secondary or supporting roles in extensive collaborations, while high values indicate active leadership in projects. In advanced career stages, this index is expected to remain above 0,5 to reflect research independence. Complementarily, the originality index measures the diversity of cited sources; a high value indicates greater interdisciplinarity. An alternative to this, for the diversity of cited sources (interdisciplinarity): The Rao-Stirling Index is an established metric for measuring diversity and interdisciplinarity based on cited references.[5]

Composite and Normalized Indicators integrate multiple dimensions of scientific activity into comprehensive measures that enable fairer, more nuanced comparisons among researchers, institutions, or countries, while accounting for the particularities of each field. Field-Weighted Citation Impact (FWCI) is a key indicator that compares the citations received by a set of publications against the global average in their respective specific fields. A value of 1 represents exactly the expected impact for the global average. Values below 1 indicate an impact below expectations, while values above 1 indicate an impact above expectations.[6]

Excellence indicators identify the percentage of a researcher›s publications that are among the top 10 % most cited worldwide in their respective years and fields. This is a high-impact threshold. Low values, for example, below 5 %, suggest that contributions are in the average range. Conversely, high values, above 10 % or especially 15 %, indicate consistently influential and elite production, demonstrating a recurring ability to generate state-of-the-art research in their field. These indicators are handy for identifying exceptional performance.[7]

New emerging indicators broaden the traditional metric spectrum into complementary dimensions, capturing aspects such as immediate visibility, the sustainability of impact, and the diversity of the research footprint. The i10 index, popularized by Google Scholar, counts the number of publications with at least 10 citations. It provides an immediate measure of the influence of moderately influential works. A low value may indicate a predominance of very recent publications that have not yet accumulated citations, or a specialization in niche fields. A high value suggests several works that have achieved significant penetration in the literature.

The impact sustainability index assesses the temporal persistence of citations received by analyzing the decay curve of citations over time. Sharply declining values may indicate early obsolescence or that the research responds to passing trends. Stable or slowly declining values, on the other hand, indicate prolonged relevance and that the contributions remain helpful within the scientific community several years after publication, a sign of fundamental work. This index is crucial for distinguishing between transient and lasting impact.

The thematic diversity index is another indicator in this group that measures the variety of scientific areas in which a researcher publishes. Low values reflect high thematic specialization, with a deep focus on a well-defined field. High values indicate research versatility and an ability to contribute to multiple disciplines, which can be an advantage in interdisciplinary environments and is increasingly valued in solving complex problems that require integrative perspectives. A researcher with a diversified profile typically has an index above 0,5 on thematic entropy scales.

The responsible interpretation of any bibliometric index requires an understanding of its methodological foundations, technical limitations, and specific disciplinary context. No single indicator captures the multidimensionality of scientific impact, so its combined and critical use is essential for balanced evaluations.

Transparency in calculation methods, the selection of appropriate indicators for each evaluative purpose, and the consideration of complementary qualitative factors are crucial principles in the ethical application of bibliometrics to contemporary scientific evaluation. Each index should be understood as a particular lens that reveals specific aspects of the complex phenomenon of scientific communication, where its value emerges from the integrated, contextualized analysis of multiple metric perspectives.

5.2. Temporal Evolution of Bibliometric Indicators

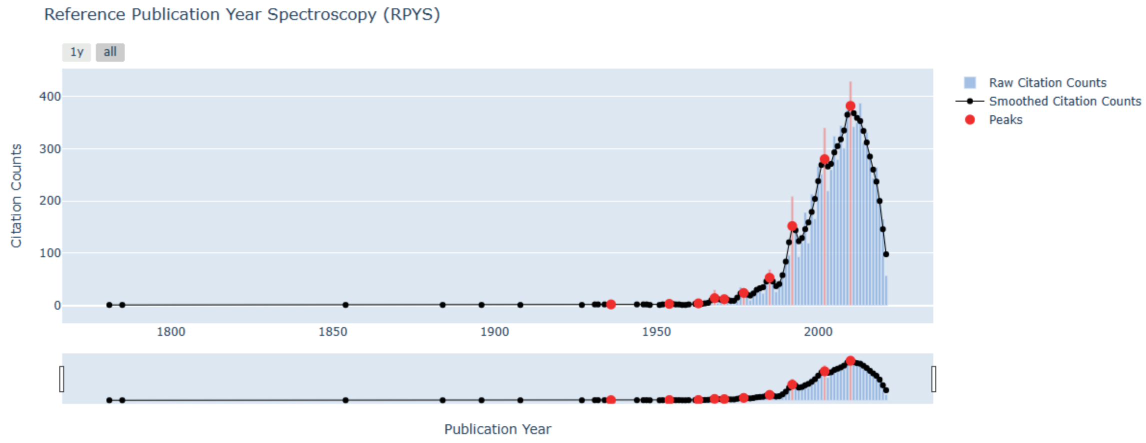


Figure 5.1. Example of a graph of citations/publications per year in spectroscopy

The temporal evolution of bibliometric indicators provides a dynamic perspective that is crucial for evaluating research trajectories, transcending the mere static snapshot offered by point values. Longitudinal analysis reveals patterns of professional development, publication strategies, and the sustained impact of scientific contributions. The interpretation of this evolution varies substantially across indicators, requiring careful contextualization within the researcher's career stage and the norms of their discipline.

In production indicators, an upward trend in the number of annual publications suggests expanding productivity, associated with the consolidation of a research group or the acquisition of funded projects. Conversely, a sustained decline may indicate a transition to more administrative roles, a deliberate strategy of quality over quantity, or difficulties in maintaining competitive activity. The evolution of the collaboration index, when it shows a progressive increase, usually reflects growing integration into broader and more complex scientific networks.

The trajectory of the h-index is particularly informative. Linear or accelerated growth in the early years of a career indicates successful scientific consolidation. During the mid-career stage, sustained growth is expected, while stabilization in the later stages may be natural. However, premature stagnation or decline could suggest a loss of scientific relevance. The m-index, when adjusted for years of activity, allows us to determine whether a researcher maintains a constant rate of impactful output or whether this rate dilutes over time.

In impact indicators, the evolution of the Field-Weighted Citation Impact (FWCI) shows how a researcher's work is received within their field. An upward trend indicates growing influence and that their recent work is receiving greater interest than their previous work. A downward trend, especially if it falls below 1, suggests that the research is not at the forefront of the field. Time windows should analyze the percentage of publications in the top 10 % of citations to identify whether excellence is being maintained, improving, or declining.

The sustainability of impact is visualized by the curve of cumulative citations over time for different publications. Seminal works exhibit a steady, steep growth curve, indicating prolonged relevance. Current works proliferate but may soon saturate, indicating temporary interest. A flat citation profile suggests limited influence. Analysis of the evolution of international collaboration, which shows an increase in its proportion, points to a progressive internationalization of the research profile, a factor highly valued in global evaluation contexts.

5.3. Combination of Bibliometric Indicators with Contextual Variables

The analytical power of bibliometric indicators is multiplied when they are systematically combined with contextual variables. These integrations transcend one-dimensional evaluation, enabling scientific activity to be dissected to answer complex questions about equity, mobility, specialization, and knowledge transfer. Each combination pursues a specific analytical goal, revealing structural patterns and underlying dynamics that would otherwise remain hidden in aggregate averages.

The combination of the journal's impact factor and the author's position and gender is used to identify potential gender biases in scientific communication. This triangulation allows us to investigate whether there are systematic differences in visibility and leadership, for example, whether male authors tend to publish as first or last authors in more prestigious journals more frequently than their female colleagues. A recurring finding across some fields is the overrepresentation of men as corresponding authors, suggesting barriers to access to positions of intellectual leadership. This approach is essential for designing evidence-based policies that promote equity in science.

The relationship between country of affiliation, international collaboration index, and normalized impact (FWCI) is used to assess the geostrategic position of national science and technology systems. A country with a high rate of international collaboration but a moderate FWCI could indicate a successful internationalization strategy that does not yet translate into global scientific leadership, suggesting a role more as a participant than as a driver. Conversely, a nation with an FWCI well above one and selective international collaboration indicates a position of scientific leadership and autonomy.

This combination is crucial for funding agencies seeking to calibrate their international cooperation policies.

Integrating researcher seniority (years since first publication) with the leadership index and thematic diversity allows for the study of career trajectories and the evolution of scientific interests. High seniority with a low leadership index may indicate a career spent mainly in support roles, while the same seniority with increasing thematic diversity suggests an evolution towards interdisciplinary interests.

This combination helps institutions understand and support different trajectories and patterns of intellectual mobility within their research staff.

The combination of document type (article, review, patent, conference proceedings) with the industrial collaboration index and citations received in patents sheds light on the mechanisms of knowledge transfer and innovation.

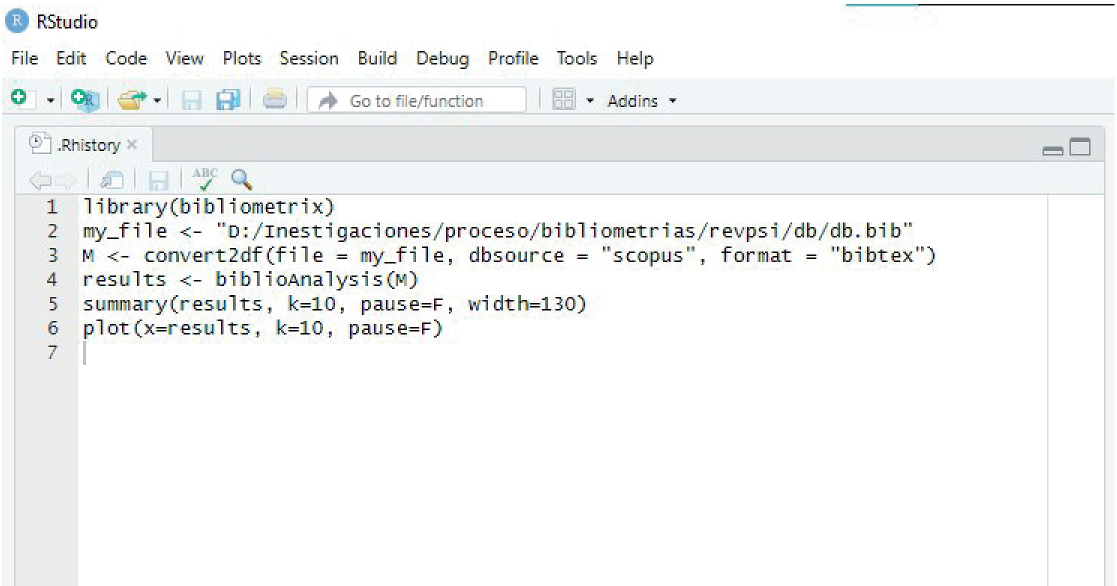
A high percentage of publications and conference proceedings, together with high industrial collaboration, suggests research oriented towards development and immediate application. In contrast, a profile dominated by basic research articles and reviews, even with high academic impact, may indicate a disconnect with productive sectors. This composite metric is vital for innovation policies.

Cross-referencing thematic specialization indicators with institutional origin (university, public research organization, hospital, company) and the centrality index in co-authorship networks allows for the mapping of knowledge ecosystems. An institution with a high specialization index and low centrality acts as a specialized but potentially isolated node. The same institution with high centrality is configured as an essential hub in that niche. These strategic combinations are indispensable for planning and informed decision-making in science policy at the macro and meso levels.

5.4. Practical calculation

5.4.1. With R/Bibliometrix (biblioAnalysis())

Calculating bibliometric indices with R and the Bibliometrix package is among the most robust and comprehensive methodologies currently available. The biblioAnalysis() function is at the core of the analytical process, processing complete bibliographic datasets to generate an object containing more than 30 different indicators. The practical implementation begins by loading the dataset into a data frame, as seen in previous sections, and then applying the primary function*: results <- biblioAnalysis(dataframe)*. This operation automatically performs all the necessary calculations, from fundamental productivity indicators to advanced collaboration and impact metrics, and returns the result as plain text.

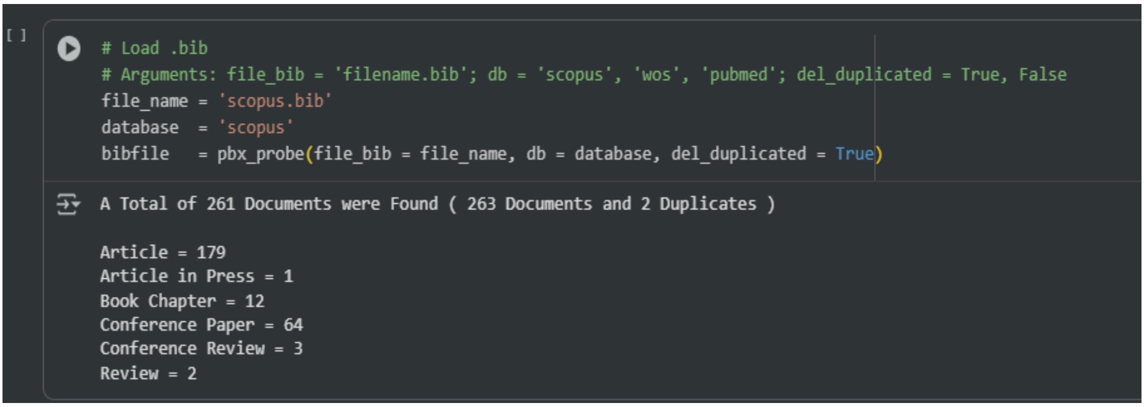

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
.Rhistory

```
library(bibliometrix)
my_file <- "D:/Inestigaciones/proceso/bibliometrias/revpsi/db/db.bib"
M <- convert2df(file = my_file, dbsource = "scopus", format = "bibtex")
results <- biblioAnalysis(M)
summary(results, k=10, pause=F, width=130)
plot(x=results, k=10, pause=F)
```

Figure 5.2. Basic project in R Bibliometrix

Specific results are extracted using complementary functions that access the object generated by *biblioAnalysis()*. To obtain the h-index and variants, *summary(results)$h.index* is used, while international collaboration statistics are retrieved with *summary(results)$Collaboration*. The *plot(results)* function generates immediate visualizations of the temporal distributions of publications and citations, which can be viewed in the lower-right panel of RStudio.

The customization of calculations in Bibliometrix allows analyses to be tailored to specific research needs. Using parameters such as sep = ";" in the initial function, the author and institution separators are adjusted based on the dataset's characteristics. For comparative analyses across periods, the timeslice function divides the time series into specific segments and calculates evolutionary indicators. Integration with other R libraries, such as igraph, enriches analytical capabilities, enabling the calculation of customized centrality indices and the analysis of communities in complex co-authorship and citation networks.



```
# Load .bib
# Arguments: file_bib = 'filename.bib'; db = 'scopus', 'wos', 'pubmed'; del_duplicated = True, False
file_name = 'scopus.bib'
database  = 'scopus'
bibfile   = pbx_probe(file_bib = file_name, db = database, del_duplicated = True)
```

```
A Total of 261 Documents were Found ( 263 Documents and 2 Duplicates )

Article = 179
Article in Press = 1
Book Chapter = 12
Conference Paper = 64
Conference Review = 3
Review = 2
```

Figure 5.3. Loading a dataset in PyBibx

The calculation of bibliometric indices using PyBibX represents a modern approach that combines the flexibility of Python with algorithms specialized in scientific literature analysis. Installation is performed with pip install pybibx, creating an environment ready to process datasets in multiple formats, including BibTeX, RIS, and CSV. Initializing the analysis requires creating a specific parser object based on the input format: *from pybibx import BibtexParser* for BibTeX files or *RisParser* for RIS formats, then loading the dataset using *parser.parse_file('file_path') or pbx_probe(), as shown in the image*. This structured approach ensures accurate interpretation of bibliographic metadata before metric calculation.

The analytical core of PyBibX resides in the MetricsCalculator class, which implements specialized methods for each indicator category. After loading the data, the calculator is instantiated using *calculator = MetricsCalculator(parsed_data)* and specific techniques are invoked, such as *calculator.h_index()* for the h-index, *calculator.g_index()* for the g-index, and *calculator.m_index()* for the time-normalized version. For collaboration analysis, *a calculator.collaboration_metrics()* generates multi-level co-authorship indicators, while *calculator.citation_analysis()* provides detailed statistics on citation distribution. Each method returns not only the numerical value but also contextual metadata that facilitates the interpretation of results.

PyBibX's advanced capabilities include temporal analysis using the *temporal_analysis()* function, which segments data by user-defined periods and calculates the evolution of indicators. To identify seminal works, *burst_detection()* applies citation burst detection algorithms. Integration with pandas allows results to be converted into DataFrames for further analysis, while native visualization utilities generate time-series graphs and collaboration networks. This combination of metric and visualization capabilities positions PyBibX as an exceptionally versatile tool for researchers who require comprehensive bibliometric analysis within the Python ecosystem.

The code example available in the official PyBibX documentation contains the code needed to run all these indexes, so, as mentioned above, it will not be necessary to memorize or write them manually.

5.4.2. Alternatives: Publish or Perish and Excel (formulas)

For researchers who require immediate solutions without programming, Publish or Perish is an alternative. Once the data is loaded, it will display various indices, such as h, g, and m, in a panel on the right.

Although Excel spreadsheets are not initially designed for bibliometric analysis, they can be used effectively through the strategic application of formulas and functions. The h-index can be calculated by sorting the citations per article in descending order and applying the formula *=MAX(ROW(A1:A100)*(A1:A100>=ROW(A1:A100)))* as a matrix, where A1:A100 contains the number of citations per publication. For the g-index, the approximation requires *=MAX(ROW(A1:A100)*(A1:A100^2>=SUM(A1:A100)))*, while the i10 index is simply calculated with *=COUNTIF(A1:A100;">=10")*. These implementations, although basic, provide independent verification of results obtained using specialized tools.

Excel's analytical capabilities are significantly enhanced by combining statistical functions with pivot tables. For collaborative analysis, the *COUNTIF* and *IF* functions allow the calculation of international and institutional co-authorship percentages using formulas such as *=COUNTIF(affiliation_range;"*country*")/COUNT.A(affiliation_range)*. Pivot tables facilitate aggregations by author, institution, or time period, while integrated charts provide immediate visualizations of citation distributions and productivity patterns. For advanced users, VBA macros enable automation of recurring calculations and the generation of standardized reports, transforming Excel into a surprisingly competent bibliometric tool for moderate-scale projects.

5.5. Critical interpretation

The interpretation of bibliometric indices must fundamentally recognize the profound differences in publication and citation cultures between academic disciplines. In hard sciences such as physics or biomedicine, publication cycles are rapid, with a high prevalence of multiple authorship and high citation rates that reflect both genuine impact and established practices of routine citation. An h-index of 20 in particle physics might be considered moderate, while the same value in the humanities would represent exceptional influence. This disparity arises from structural differences: the sciences generate more publications per researcher per year and have developed citation traditions that prioritize the constant updating of references.

The humanities and social sciences operate under radically different paradigms, in which books constitute the primary format for scientific communication and evaluation cycles are substantially longer. The specialized monograph, with its deep and sustained argumentation, rarely receives metric recognition proportional to its actual intellectual influence when indicators designed for scientific journal articles are applied.[8]

Engineering and applied technologies present hybrid patterns in which academic citations coexist with other indicators of impact, such as patents, technological developments, and transfers to the productive sector. An engineering researcher may have a modest h-index while generating innovations with significant industrial impact, creating a dangerous disconnect between academic metrics and real impact. These disciplines often publish in specialized conferences whose citations are not fully captured by traditional bibliographic databases, systematically underestimating their intellectual implications.

The health sciences show particular biases derived from the concentration of citations in systematic reviews and meta-analyses, which receive disproportionately more citations than primary studies. A clinical researcher engaged in complex longitudinal studies but with limited samples may appear less influential than colleagues who publish frequent reviews, distorting the evaluation of substantive contributions to the advancement of medical knowledge. Extreme specialization across subfields with varying sizes of academic communities introduces additional variation that invalidates direct comparisons.

Interdisciplinary evaluation represents the most complex challenge, where researchers working at the frontiers between fields face a double penalty: their publications may be less cited in each discipline while remaining outside the core areas that concentrate citations. A computational neuroscientist, for example, might publish in neuroscience and computer science journals, dividing their impact between two communities with different citation practices and failing to reach critical visibility thresholds in either one separately, despite potentially transformative contributions at the disciplinary intersection.

Disciplinary normalization attempts to correct these biases through comparisons within defined fields, but faces significant methodological limitations. Indicators such as Field-Weighted Citation Impact (FWCI) use broad categories that often group subdisciplines with heterogeneous publication cultures. An analytical philosopher and a historian of philosophy share a category but operate in substantially different citation ecosystems, effectively invalidating many statistical corrections. The solution lies in supplementing quantitative metrics with contextualized qualitative assessment by disciplinary experts who understand these subtle but crucial differences.

Recap

- Bibliometric indices are quantitative measures used to evaluate scientific productivity, impact, and influence.
- They serve as a basis for comparing authors, journals, institutions, or countries within the same field of knowledge.
- They are mainly derived from publication and citation records in databases such as Web of Science, Scopus, and Google Scholar.
- The indicators are grouped into three broad categories:
    - Scientific productivity (number of publications).
    - Impact or influence (citations received).
    - Collaboration and networks (co-authorships, affiliations).

- The Impact Factor (IF), created by Eugene Garfield, measures the average number of citations received by articles in a journal over two years.
- CiteScore (Elsevier) evaluates citations received over four years, covering more journals than the IF.
- The SCImago Journal Rank (SJR) weights citations according to the relevance of the issuing journals, based on the PageRank algorithm.
- The Source Normalized Impact per Paper (SNIP) adjusts citation impact to account for differences across disciplines.
- At the author level, the most commonly used indicators are the h-index and the g-index.
- The h-index reflects the balance between productivity and citation: an author has an h-index if they have published h articles with at least h citations each.
- The g-index, proposed by Egghe, gives greater weight to the most cited articles.
- There are variants such as h5, hm, hg, hI, and hIa, which are applied to adjust for differences in academic age or co-authorship.
- For scientific journals, the most common indicators include FI, SJR, CiteScore, SNIP, and Eigenfactor.
- The Eigenfactor measures a journal’s importance by considering the entire structure of its citation network.
- Collaboration indices are calculated using co-authorship, institutional networks, and affiliation analysis.
- In institutional or national evaluation, aggregate metrics are used (e.g., total number of citations or publications per field).
- Indicators should be interpreted in context, avoiding comparisons between disciplines with different citation habits.
- The use of standardized indicators and complementary metrics (altmetrics, downloads, social visibility) is recommended.
- Bibliometric indices, although useful, have ethical and methodological limitations when used as the sole evaluation criteria.
- A responsible evaluation combines quantitative indicators and qualitative review, in accordance with the San Francisco Declaration on Research Assessment (DORA) and the Leiden Manifesto.

Self-assessment questions

1. What is the primary function of bibliometric indices?
2. Into which three broad categories are bibliometric indicators grouped?
3. Who created the Impact Factor, and what is its fundamental principle?
4. How does CiteScore differ from the Impact Factor?
5. How does the SJR indicator weight citations?
6. What does the h-index measure, and how is its value interpreted?
7. What does the g-index contribute to the h-index?
8. What characterizes the SNIP indicator in comparison with other indices?
9. What principles do DORA and the Leiden Manifesto promote in scientific evaluation?
10. Why is it necessary to contextualize the results of bibliometric indices?

BIBLIOGRAPHY

1. Moed HF. Citation analysis in research evaluation. Dordrecht: Springer; 2005. doi: 10.1007/1-4020-3714-7.

2. Bornmann L, Daniel HD. What do citation counts measure? A review of studies on citing behavior. J Doc. 2008;64(1):45–80. doi: 10.1108/00220410810844150.

3. Gingras Y. Bibliometrics and research evaluation: Uses and abuses. Cambridge (MA): MIT Press; 2016. ISBN: 9780262337663.

4. Egghe L. Theory and practice of the g-index. Scientometrics. 2006;69(1):131–52. doi: 10.1007/s11192-006-0144-7.

5. Leydesdorff L, Bornmann L, Mutz R, Opthof T. Turning the tables on citation analysis one more time: Principles for comparing sets of documents. J Informetrics. 2019;13(3):1020–30. doi: 10.1016/j.joi.2019.05.010.

BIBLIOGRAPHIC REFERENCES

1. Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151–5. https://doi.org/10.1087/20150211

2. Bornmann L, Bauer J. Evaluation of the highly-cited researchers' database for a country: proposals for meaningful analyses on the example of Germany. Scientometrics. 2015;105(3):1997–2003. https://doi.org/10.1007/s11192-015-1754-3

3. Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences. 2005;102(46):16569–72. https://doi.org/10.1073/pnas.0507655102

4. Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AFJ. Towards a new crown indicator: some theoretical considerations. Journal of Informetrics. 2011;5(1):37–47. https://doi.org/10.1016/j.joi.2010.08.001

5. Stirling A. A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface. 2007;4(15):707–19. https://doi.org/10.1098/rsif.2007.0213

6. Purkayastha A, Palmaro E, Falk-Krzesinski HJ, Baas J. Comparison of two article-level, field-independent citation metrics: field-weighted citation impact (FWCI) and relative citation ratio (RCR). Journal of Informetrics. 2019;13(2):635–42. https://doi.org/10.1016/j.joi.2019.02.005

7. Elsevier. SciVal metric – Pure. 2025. https://helpcenter.pure.elsevier.com/en_US/metrics-in-pure/scival-metric

8. Kulczycki E, Engels TCE, Pölönen J, Bruun K, Dušková M, Guns R, et al. Publication patterns in the social sciences and humanities: evidence from eight European countries. Scientometrics. 2018;116(1):463–86. https://doi.org/10.1007/s11192-018-2711-0

Chapter 6 / Capítulo 6

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch06

Correlation Graphs / Grafos de Correlación

6.1. Theoretical foundations of bibliometric graphs

Correlation graphs are a fundamental tool for analyzing relationships in scientific literature, allowing connections between academic entities to be visualized and quantified. Based on mathematical graph theory, these structures transform bibliographic data into networks in which meaningful relationships interconnect elements of the research system. The analytical power of bibliometric graphs lies in their ability to reveal underlying structural patterns in large volumes of data, facilitating the identification of intellectual communities, emerging trends, and collaborative dynamics that would remain hidden in conventional tabular analyses.

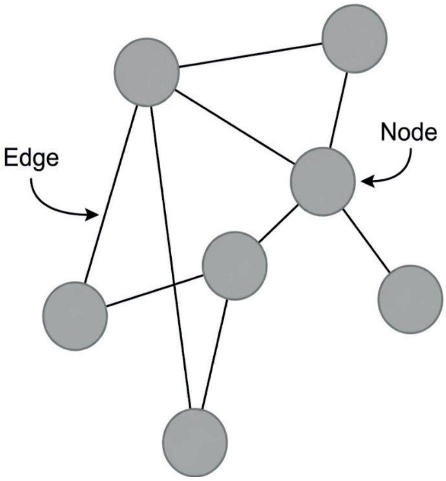


Figure 6.1. General structure of a graph

6.1.1. Nodes: authors, keywords, and journals as units of analysis

Nodes are the fundamental elements of the graph, representing the academic entities whose study reveals the structure of knowledge. When nodes represent authors, the graph visualizes collaboration communities, where relative position indicates centrality in co-authorship networks, and proximity suggests thematic or institutional affinity. Highly connected researchers function as bridges between different groups, while peripheral nodes may indicate thematic specialization or academic isolation. The density of connections around a specific author reflects their degree of integration into the scientific community, providing insights into collaboration strategies and intellectual leadership.

In keyword-based maps, nodes capture research concepts and themes, transforming intellectual content into interpretable spatial structures. The frequency of term occurrence determines node size, while co-occurrence within the same documents establishes connections between concepts. This approach reveals the conceptual architecture of a scientific field, identifying consolidated thematic nuclei, emerging areas, and research gaps. Natural terminological clusters emerge from community detection algorithms, revealing how knowledge is organized and how different subfields relate to one another within a disciplinary domain.

Scientific journals as nodes allow us to analyze the structure of academic communication and knowledge flows between different specialties. Citation graphs between journals show relationships of intellectual influence, where the direction of citations indicates knowledge flows and the strength of connections reflects disciplinary proximity. This approach identifies central journals that function as dissemination centers, bridge publications that connect different fields, and specialized peripheries. Diachronic analysis of these networks can reveal the evolution of disciplinary boundaries and the emergence of new interdisciplinary research areas.

6.1.2. Edges: weight and direction as measures of relationship

Edges are the connecting elements of the graph, quantifying relationships between nodes along two critical dimensions: weight and direction. The weight of an edge reflects the intensity of the connection and is calculated using different metrics depending on the type of analysis. In co-authorship networks, weight can represent the number of joint publications; in co-word networks, the frequency of terminological co-occurrence; in citation networks, the volume of citations between entities. This weighting allows us to distinguish between occasional connections and solid, sustained relationships, which is essential for identifying strategic collaborations and consolidated thematic nuclei.

The direction of the edges introduces a temporal and causal dimension to the analysis, which is particularly crucial in studies of citation and intellectual influence. In citation networks between authors or journals, directed edges indicate flows of knowledge, showing who cites whom and revealing patterns of academic influence. This directionality allows us to identify seminal works that receive many citations but cite few (sink nodes), versus synthesis works that cite extensively and are widely cited (hub nodes). Centrality analysis in directed networks provides more nuanced measures of influence that consider both the impact received and the capacity for knowledge dissemination.

The interaction between weight and direction adds a layer of analysis to the interpretation of bibliometric graphs. A heavy, bidirectional edge between two authors suggests intense, reciprocal collaboration, while a light, unidirectional edge may indicate incidental influence or one-off recognition. In co-citation analysis, the strength and reciprocity of connections reveal intellectual proximity and membership in shared schools of thought. Community detection in weighted and directed graphs identifies groups with high internal cohesion and characteristic patterns of external connections, which are essential for understanding the social and intellectual structure of scientific fields.

Bibliometric visualization programs incorporate distinctive features that determine the graphic significance and interpretive capacity of the graphs generated. VOSviewer, for example, uses layout algorithms based on attraction and repulsion that position nodes according to their similarity, creating natural clusters where visual distance represents thematic or collaborative proximity. CiteSpace, in contrast, uses temporal representations that show the evolution of networks through timelines, where the vertical position indicates publication time and the horizontal position reflects citation relationships. These algorithmic differences produce visualizations with different capacities to reveal specific patterns, making it essential to understand the representation principles of each tool.

The addition of clusters or groupings represents an advanced functionality that substantially enriches the interpretation of graphs. These clusters are generated using community detection algorithms that identify subgroups of nodes with high internal connectivity and lower external connectivity, typically differentiated by distinctive colors. In co-word analysis, clusters reveal consolidated research topics; in co-authorship networks, they show collaboration groups; in citation networks, they identify schools of thought or shared paradigms. Automatic clustering tagging using representative terms extracted from nodes is a key feature that facilitates immediate interpretation of the field of study's underlying structure.

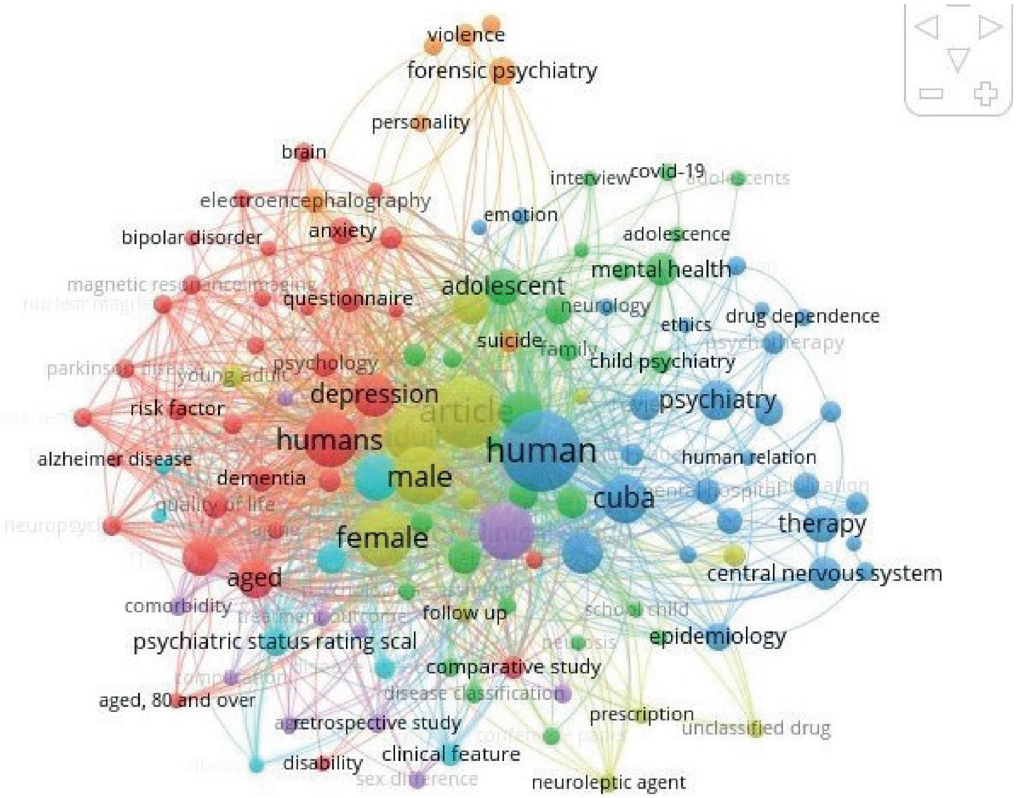


Figure 6.2. Example of a co-occurrence graph of terms or words

The most sophisticated tools incorporate additional elements such as concentric rings representing temporal dimensions, node sizes proportional to impact metrics, and edge thicknesses reflecting relationship intensity. Some programs, such as CitNetExplorer, add timelines showing the diachronic evolution of connections, while SciMAT introduces strategic maps that position topics according to their density and centrality. These layers of information transform graphs from simple structural representations into multidimensional analytical tools that simultaneously capture relationships, intensity, temporal evolution, and impact, requiring the user to understand these visual conventions to extract meaningful insights from the generated visualizations.

6.2. Key algorithms for constructing bibliometric graphs

Graph construction algorithms represent the methodological core of bibliometric network analysis, transforming bibliographic data into meaningful relational structures. Each algorithm operates under different conceptual principles and answers different research questions, requiring the analyst to have a deep understanding of its theoretical foundations and practical limitations. The appropriate selection of the algorithm determines not only the validity of the results but also the ability to extract meaningful insights into the structure and dynamics of scientific knowledge. This chapter examines in depth three fundamental algorithms—co-occurrence, bibliographic coupling, and co-citation—unraveling their operating mechanisms, characteristic applications, and optimal contexts of use.

6.2.1. Co-occurrence analysis: mapping conceptual structure

Co-occurrence analysis is based on the principle that the joint appearance of terms in academic documents reveals conceptual and thematic proximity. This algorithm constructs networks in which nodes represent concepts, typically extracted from keywords, titles, or abstracts, and edges reflect their co-occurrence within the same document. The technical implementation involves multiple processing stages: initially, the text is normalized through lemmatization or stemming to unify morphological variants; subsequently, non-significant terms are filtered using domain-specific stopword lists; finally, a document co-occurrence matrix is constructed where each cell records the frequency of joint occurrence.

Measuring the associative strength between terms requires the application of advanced normalization coefficients. The Jaccard index, calculated as the ratio of observed co-occurrences to the union of individual frequencies, is particularly effective at correcting for bias toward widespread terms. Alternatively, the cosine similarity coefficient measures the angle between frequency vectors in multidimensional spaces, providing robustness for analyzing large volumes of data. For contexts where terminological rarity contains valuable information, measures such as weighted specific association preserve connections between specialized terms. The selection of the appropriate coefficient depends critically on the terminological distribution of the analyzed corpus and the particular research objectives.

The applications of co-occurrence analysis range from identifying emerging research niches to mapping the cognitive structure of established disciplines. In technology watch, it enables the detection of convergences between previously disjoint fields, indicating potential areas of innovation. In interdisciplinary studies, it reveals conceptual bridges between different epistemological domains. The temporal evolution of co-occurrence networks, obtained through diachronic analysis segmented by periods, shows the dynamics of the formation, consolidation, and dissolution of research topics, providing unique perspectives on the processes of conceptual change in science.

6.2.2. Bibliographic coupling: connections through shared references

Bibliographic coupling establishes relationships between documents based on their shared citation profile, operating on the principle that works that cite familiar sources are likely to share theoretical, methodological, or conceptual frameworks. Unlike other algorithms, bibliographic coupling generates static networks whose connections are fixed at the time of publication and do not evolve. This feature makes it particularly valuable for analyzing recent literature that has not yet accumulated enough citations for other types of relational analysis.

The technical implementation requires the construction of a document-reference matrix where each row represents a document in the corpus and each column a cited reference. The matrix is then transformed into a similarity network through matrix multiplication and the application of association measures. The simple coupling index, based on the raw count of shared references, tends to favor documents with extensive reference lists, so in practice, normalized measures such as the Salton index are preferred, which divides the number of shared references by the square root of the product of the total references in each document. For specialized analysis, the bibliographic proximity index weights references by age, giving greater weight to recent citations.

The applications of bibliographic coupling are particularly relevant in the context of contemporary literature evaluation and the mapping of active scientific frontiers. In systematic reviews of recent advances, it allows the identification of groups of publications with similar approaches without waiting for citation traditions to be established. In interdisciplinary studies, documents that function as bridges between fields are revealed by their mixed citation patterns. The main methodological limitation lies in its sensitivity to individual citation styles and its inability to capture indirect influences or subsequent developments in intellectual relationships.

6.2.3. Co-citation analysis: the collective perception of the scientific community

Co-citation analysis is based on the principle that the joint citation of two documents by subsequent works reflects a relationship perceived by the scientific community. Unlike bibliographic coupling, which examines citations from source documents, co-citation analyzes citations from citing documents, thereby capturing a collective, dynamic assessment of intellectual relationships. This algorithm generates evolving networks in which the strength of connections between documents can increase, decrease, or reconfigure over time, reflecting changes in scientific consensus about the structure of knowledge.

The construction of co-citation networks involves complex technical phases, including the identification of co-cited pairs, the calculation of co-citation frequencies, and the application of minimum thresholds to include meaningful connections. Normalization of co-citation strength typically employs the standard co-citation index, which is simply the count of common citers. However, advanced implementations use Pearson's correlation coefficient to capture similarities in co-citation patterns over time. For diachronic analysis, the data is segmented into successive time windows, allowing the evolution of intellectual groupings and the emergence of new lines of research to be tracked.

The applications of co-citation analysis are compelling in historical studies of science and analyses of the evolution of intellectual paradigms. It allows for the identification of seminal works that have remained relevant over time, the detection of mergers between previously separate intellectual traditions, and the analysis of processes of specialization and disciplinary fragmentation. In scientific evaluation, it provides robust indicators of lasting intellectual impact beyond conventional citation metrics. Among its limitations are the inherent time lag before documents accumulate enough citations for meaningful analysis and the possible reinforcement of established canons to the detriment of marginal but potentially transformative contributions.

Author co-citation graphs

Author co-citation graphs represent the connections between researchers who are cited together in the references of other works. When two authors are frequently mentioned together in the same publications, it can be inferred that their contributions belong to a common theoretical or thematic framework. The intensity of co-citation reflects their shared influence in a field of study, allowing us to identify schools of thought, intellectual leaders, and the invisible structure of scientific trends. A temporal analysis of these graphs can reveal the evolution of paradigms and the emergence of new approaches.

Document co-citation graphs

These graphs map the relationships between publications that are simultaneously cited by other articles. Each node represents a document, and the edges symbolize its co-occurrence in the bibliographic sections of subsequent works. This structure allows us to identify the literary foundations of a field: the seminal articles that form the theoretical core, as well as the peripheral works that connect different areas. The density of connections around a document indicates its centrality in the construction of disciplinary knowledge.

Co-citation graphs of countries and institutions

At the macro level, co-citation graphs can be applied to countries or institutions, with nodes representing these entities and edges reflecting the frequency with which their research is cited together. This indicates a thematic or methodological proximity between their scientific systems. Two countries with high co-citation rates tend to specialize in similar research niches or collaborate intensively. These graphs help analyze international competitiveness, identify clusters of excellence, and design global positioning strategies in science and technology.

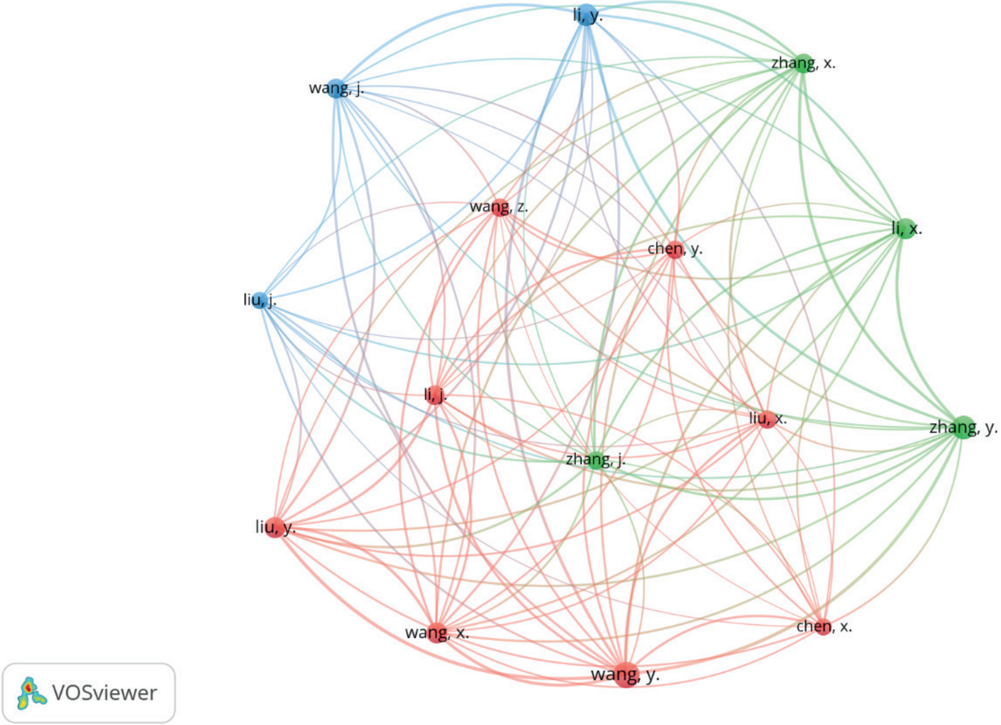


Figure 6.3. Example of a co-citation graph

6.2.4. Methodological integration and comparative perspectives

Specific epistemological, temporal, and pragmatic considerations should guide the selection between these algorithms. Co-occurrence analysis is optimal for exploring the immediate conceptual structure of a field, particularly in disciplines where specialized terminology efficiently encodes intellectual content. Bibliographic coupling offers decisive advantages for analyzing recent literature and identifying contemporary intellectual alignments. Co-citation analysis, on the other hand, provides historical depth and captures consolidated collective assessments of intellectual relationships.

The most sophisticated methodological approaches combine multiple algorithms in triangulated designs that mitigate the individual limitations of each method. A typical design might employ co-occurrence to identify emerging themes, bibliographic coupling to analyze their current structuring, and co-citation to contextualize their historical development. This multi-method integration allows us to distinguish between immediate conceptual connections (co-occurrence), contemporary intellectual alignments (bibliographic coupling), and consolidated perceptions of relationships (co-citation), thereby producing richer, methodologically robust bibliometric analyses.

A deep understanding of these algorithms, their theoretical foundations, practical implementations, and characteristic biases is an essential skill for any researcher who aspires to produce rigorous and meaningful bibliometric analyses. Far from interchangeable tools, each algorithm illuminates distinct dimensions of the complex topography of scientific knowledge, requiring conscious selection and contextualized application to realize its analytical potential fully.

6.3. Visual tools

The specialized tools used to construct graphs in the field of bibliometrics include VOSviewer, CiteNetExplorer, CiteSpace, R Bibliometrix, and Pybibx. However, the latter two are not specialized in graph construction. The procedure for each is as follows:

6.3.1. VOSviewer: creation of thematic maps

1. Export complete bibliographic data from Scopus, WoS, or PubMed in compatible formats (RIS, CSV, EndNote).
2. Start VOSviewer and select “Create” → “Create a map based on bibliographic data.”
3. Choose the type of analysis according to the research objective:
   - Co-occurrence for analysis of terms and concepts.
   - Co-authorship for scientific collaboration networks.
   - Co-citation for citation-based intellectual structures.
   - Bibliographic coupling for relationships through shared references.

4. Configure threshold parameters according to data volume:
   - Minimum number of occurrences of a term (typically 5-10).
   - Minimum number of documents per author (2-5 for collaboration analysis).
   - Minimum number of citations per reference (10-20 for co-citation).

5. Apply the VOS clustering and visualization algorithm.
6. Customize the final visualization:
   - Adjust node size according to frequency or impact metrics.
   - Modify the color scheme to differentiate clusters.
   - Configure labels and zoom levels to optimize readability.
   - Apply network smoothing to reduce visual overlap.

The fundamental feature of VOSviewer lies in its visualization algorithm, which is based on multidimensional stress function minimization techniques, where the distance between nodes directly represents their similarity, calculated using normalized association measures. This tool generates maps where clusters emerge naturally as dense spatial groupings, differentiated chromatically to facilitate immediate identification. The visual representation prioritizes thematic interpretability over absolute metric precision, making VOSviewer the preferred choice for exploratory analysis and communication of results to non-specialist audiences.

6.3.2. CitNetExplorer: citation map creation

1. Prepare complete citation data from Web of Science in standard export format
2. Load the data file via “File” → “Open Citation Network.”
3. Configure the time and impact filtering parameters:
   - Define the range of years to be analyzed.
   - Set the minimum number of citations for document inclusion.
   - Specify selection criteria for seminal documents.

4. Generate the complete citation network with all interconnections.
5. Apply specific exploration tools:
   - Use the zoom function to examine particular time periods.
   - Apply dynamic filters by number of citations or year of publication.
   - Use the connected components selection tool.

6. Perform citation trajectory analysis:
   - Identify foundational documents and their lines of descent.
   - Trace paths of intellectual development between publications.
   - Analyze patterns of thematic convergence and divergence.

CitNetExplorer stands out for its explicit temporal representation, in which the vertical axis unambiguously encodes publication years, creating a visual timeline that reveals the diachronic evolution of citation relationships. This specialized tool allows you to trace the historical development of scientific ideas through citation connections between publications, showing how concepts are transmitted, transformed, and branched over time. The ability to animate the evolution of the network year by year provides unique insights into patterns of intellectual inheritance and moments of paradigmatic change in scientific development.

6.3.3. CiteSpace: Detection of Research "Bursts."

1. Download and install CiteSpace with the updated Java Runtime Environment.
2. Create a new project and configure the directory structure for data and results.
3. Import and convert bibliographic data from WoS or Scopus to CiteSpace's native format.
4. Configure the temporal analysis parameters exhaustively:
   - Divide the study period into time segments (1-3 years recommended).
   - Establish selection criteria by percentile (Top 10 %, 20 %, etc.) or by top N per segment.
   - Define network pruning strategies (Pathfinder, Minimum Spanning Tree, or none).

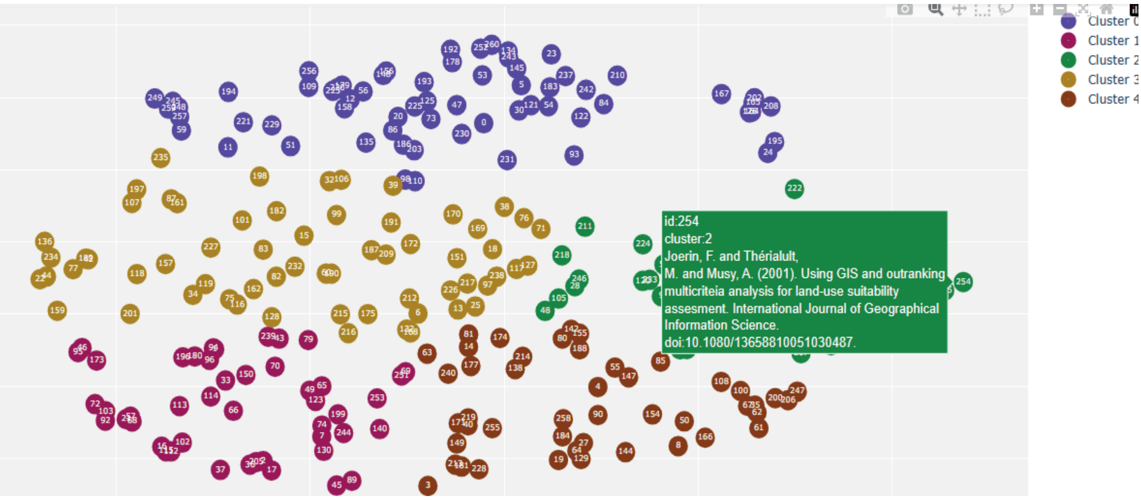
5. Run the burst detection algorithm using:
   - Select entities for analysis (terms, references, authors).
   - Configuring Kleinberg algorithm parameters for detection sensitivity.
   - Specifying minimum thresholds for burst duration and intensity.

6. Generate and analyze integrated visualizations:
   - Interpret concentric rings of annual citations in nodes.
   - Identify thematic clusters through automatic tag analysis.
   - Analyze centrality metrics and connections for bridge nodes.

The uniqueness of CiteSpace lies in its ability to integrate multiple analytical dimensions into a single visualization: spatial position indicates similarity relationships, concentric rings show temporal citation patterns, colors differentiate thematic clusters, and red rings highlight burst periods of citation activity. This tool uses specialized algorithms to detect turning points in scientific literature, identifying not only emerging topics but also specific moments of acceleration in research attention. The resulting visual representation provides a dynamic map of scientific evolution that simultaneously captures spatial structure, temporal development, and intellectual turning points.

6.3.4. Graphs with PyBibx

Both R Bibliometrix and Pybibx offer functionality for constructing co-authorship graphs, each with its own algorithmic details. Pybibx is characterized by implementing a more straightforward construction procedure with flexible variants. One of these representations generates a graph in which nodes representing individual articles, referenced by their unique IDs, interact without explicit edges. This alternative visualization enables analysis of the proximity or coexistence of publications within a conceptual or temporal space defined by the researcher, offering a different perspective from that of a traditional network of direct links.



id:254
cluster:2
Joerin, F. and Thérialult,
M. and Musy, A. (2001). Using GIS and outranking multicriteia analysis for land-use suitability assesment. International Journal of Geographical Information Science.
doi:10.1080/13658810051030487.

Figure 6.4. Example of a co-occurrence graph of terms or words

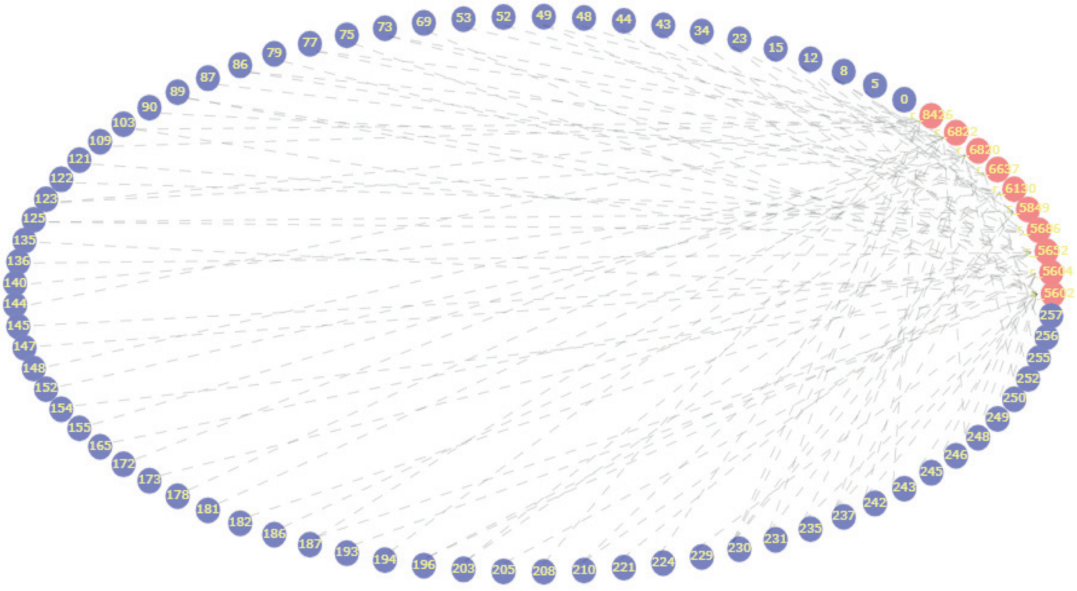


Figure 6.5. Example of a co-occurrence graph of terms or words

Beyond edgeless representations, these tools also allow the generation of classic relational graphs. In these graphs, the connections or edges between nodes are explicitly defined, representing citation or co-citation links between individual articles. Unlike maps that group elements into thematic clusters, this visualization focuses on showing the network of direct connections and its structure in its purest form, without applying clustering algorithms. This allows the analyst to identify connection patterns, bridge articles, or network density without the influence of prior automatic clustering.

One of the most eloquent visualizations of the spatial dimension of research is a graph superimposed on a geopolitical map. In this representation, nodes are positioned at the geographic locations of the authors' affiliations, illustrating the global distribution of scientific production. The edges connecting these institutions represent co-authorship links, allowing for immediate visualization of the flows and intensity of international collaboration. This graph is indispensable for identifying centers of scientific gravity, patterns of regional cooperation, and the geographical projection of research networks, offering an extremely valuable layer of contextual analysis.

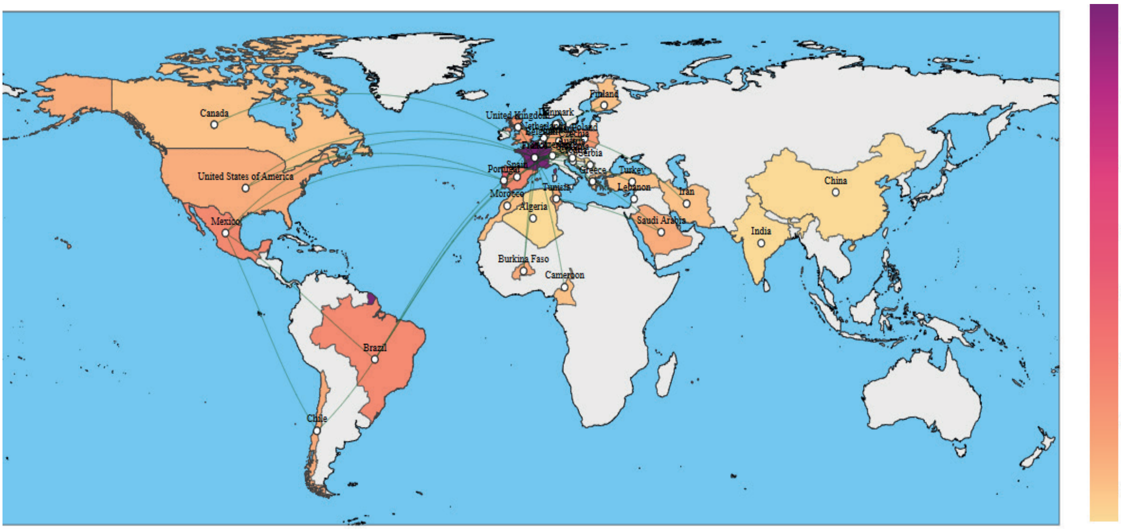


Figure 6.5. Example of a co-citation graph with the geographical distribution of authors

PyBibx offers other graph types that can be consulted in its official documentation and examples.

6.3.5. Graphs with R Bibliometrix

R Bibliometrix can create graphs of author co-citations, bibliographic links between documents, and collaboration between countries and institutions based on co-authorship relationships. To create graphs in R Bibliometrix, proceed as follows:

1. Install packages:

install.packages("bibliometrix")
install.packages("tidyverse")

2. Load libraries

library(bibliometrix)
library(tidyverse)

3. Load and process data:

```
File <- “file.bib”
M <- convert2df(file, dbsource = “wos”, format = “bibtex”)
```

4. Create a co-occurrence network:

```
NetMatrix <- biblioNetwork(M, analysis = “co-occurrence”,
                     network = “author_keywords”,
                     sep = “;”)
```

5. Create graph:

```
net <- networkPlot(NetMatrix,
              n = 50,  # Number of terms to display
              type = «fruchterman”,  # Layout
             Title = «Term Co-occurrence - Author Keywords»,
              labelsize = 0,8,
              size = 5,
              remove.isolates = TRUE,
              cluster = «walktrap”)  # Clustering method
```

All of these are displayed in the plot panel or can be exported to an image using:

```
png(“term_co-occurrence.png”, width = 1200, height = 900, res = 150)
print(net)
```

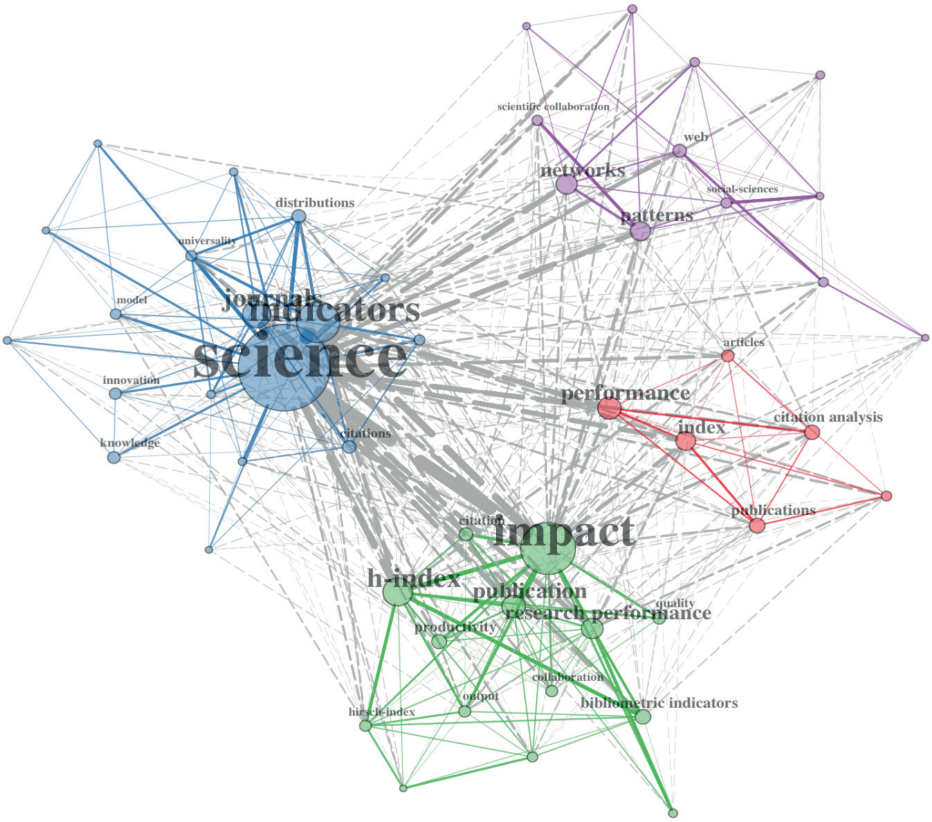


Figure 6.6. Example of a co-occurrence graph

6.4. Network interpretation

6.4.1. Identification of thematic clusters

The identification of thematic clusters represents the analytical process by which conceptual communities within a bibliometric network are discovered and delimited. These clusters emerge naturally from community detection algorithms that identify subgroups of nodes with high internal connectivity and relatively low external connectivity. In practice, each cluster encapsulates a coherent thematic domain, where nodes, whether terms, authors, or publications, share significant semantic, methodological, or theoretical characteristics. The interpretation of these clusters requires a multidimensional analysis that considers not only the internal composition of the group but also its relationships with other clusters and its position within the overall network structure.

The interpretation process begins with an examination of the representative labels that algorithms automatically assign to each cluster, typically derived from the most frequent or central terms within each group. However, this automatic approach must be complemented by a qualitative assessment that examines representative publications from each cluster to understand their substantive content.

The validation of internal thematic coherence is carried out by analyzing the foundational documents, those with the most extraordinary centrality within the cluster, and identifying the core concepts that define the thematic identity of the group. This dual quantitative and qualitative approach allows us to transcend mere structural description to achieve a deep understanding of the intellectual meanings encapsulated in each grouping.

The relative position of clusters within the global map provides crucial information about the structure of the field of study.

Spatially proximate clusters typically share conceptual frameworks, methodologies, or applications, whereas distant clusters represent distinct intellectual traditions or specialized fields. The connections between clusters, manifested as bridge links, point to potentially fertile thematic interfaces for interdisciplinary research. Diachronic analysis of the evolution of these clusters reveals dynamics of disciplinary fragmentation, paradigm fusion, or the emergence of new hybrid fields, providing valuable insights into the trajectory of knowledge development in the domain under study.

6.4.2. Centrality vs. density

Centrality and density are two fundamental analytical dimensions that capture complementary aspects of the structure and dynamics of bibliometric networks. Centrality measures the strategic position of a node within the global network, identifying elements that function as connectors between different regions of the graph. In contrast, density quantifies the degree of internal interconnection within a specific cluster or subnetwork, reflecting the cohesion and mature development of a thematic community. The joint interpretation of these metrics allows for a sophisticated characterization of the field of study's intellectual architecture and the specific roles that different elements play within that cognitive ecosystem.

In analytical practice, centrality manifests itself in multiple forms, each revealing different aspects of structural influence.

Degree centrality identifies nodes with many direct connections, typically fundamental concepts or highly collaborative authors. Intermediation points to nodes that connect different clusters, functioning as conceptual bridges between separate thematic communities. Closeness detects nodes that can quickly reach the rest of the network, indicating concepts or authors with broad diffuse influence. For example, in a co-word map of artificial intelligence, terms such as “machine learning” would show high centrality.

At the same time, “explainable AI” could exhibit high intermediation by connecting clusters of computational ethics and learning algorithms.

Density, on the other hand, characterizes the internal development of thematic clusters. High-density clusters, with numerous internal connections, represent mature, highly structured fields in which the constituent concepts have well-defined, stable relationships. Low-density clusters suggest emerging or developing areas, where conceptual relationships are still incipient, and the internal structure is still forming. In an authorship analysis, a dense cluster would indicate a consolidated collaboration group with multiple joint projects.

In contrast, a sparse cluster would indicate an emerging collaboration network with more sporadic or bilateral interactions.

The combination of these dimensions in strategic matrices, such as centrality versus density analysis, provides a robust framework for the typological characterization of clusters and nodes. For example, driving themes appear as clusters with high density and centrality, representing consolidated areas that drive the field’s development. Niche topics show high density but low centrality, indicating mature but isolated specializations.

Emerging topics exhibit low density but high centrality, signaling bridge concepts with potential for future development. Peripheral topics have low density and low centrality, representing incipient specializations or areas in decline. This typology enables strategic prioritization of research areas based on specific scientific development or research policy objectives.

In this interpretation phase, technical mastery of bibliometrics is necessary but not sufficient to extract substantive meaning from the networks generated. The researcher must have in-depth knowledge of the specific domain of study to establish meaningful connections between seemingly disparate terms and recognize conceptual relationships that transcend mere statistical coincidences. This specialized disciplinary understanding allows one to discern between superficial terminological associations and deep intellectual links, between passing terminological fads and foundational concepts with actual structuring capacity. Thematic expertise thus becomes the indispensable lens through which quantitative patterns acquire qualitative meaning and intellectual relevance.

The capacity for abstraction emerges as a critical competence for transcending immediate visualization and constructing mental models that explain the underlying architecture of the knowledge represented.

This ability allows fragmentary findings to be integrated into coherent narratives about the structure and dynamics of the field of study, identifying not only what the networks explicitly show but also what they implicitly suggest through their gaps, asymmetries, and relational patterns. The researcher must constantly move between the micro level of individual terms and specific connections, the meso level of thematic clusters and their interrelationships, and the macro level of the field’s overall structure, synthesizing perspectives across multiple scales of analysis into an integrated, hierarchically organized understanding.

The final interpretation, therefore, represents a creative synthesis where metric rigor is combined with disciplinary sensitivity and the researcher's inferential capacity. This process transforms relational data into actionable knowledge, identifying research opportunities, revealing fertile interfaces between specialties, and proposing explanatory narratives about the field's evolution and current state. The quality of this interpretation depends fundamentally on the researcher's ability to exercise informed judgment, contextualize findings within broader intellectual traditions, and communicate complex perspectives in an accessible manner without sacrificing analytical depth or conceptual precision.

An example of a graph and its interpretation is as follows:

Context: search result in SCOPUS for: *(TITLE-ABS-KEY (cardiolog* OR cardiac OR heart OR coronary OR "myocardial infarction" OR arrhythmia OR echocardiography OR hypertension)) AND (AFFILCOUNTRY (Cuba)*

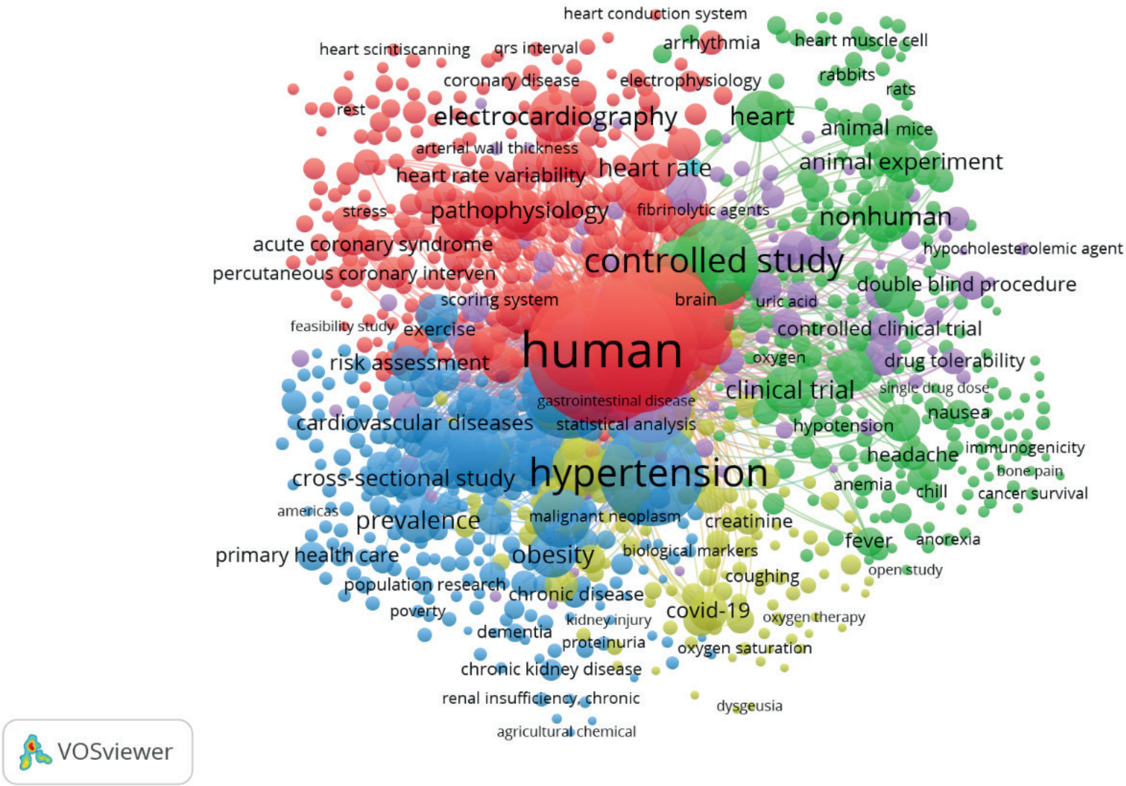


Figure 6.6. Example of a co-occurrence graph

Possible interpretation

Red cluster (top left): this cluster focuses on clinical cardiology and electrophysiology. It includes terms such as "heart," "electrocardiography," "arrhythmia," "heart rate," "coronary disease," and "acute coronary syndrome." This shows a focus on research into coronary heart disease and arrhythmias, and on the use of diagnostic techniques such as electrocardiograms. The connections to "human" and "controlled study" demonstrate that these studies are conducted on patients.

Blue cluster (bottom left): this cluster focuses on the epidemiology and risk factors of cardiovascular diseases. Key terms are “hypertension,” “cardiovascular diseases,” “prevalence,” “cross-sectional study,” “primary health care,” and “obesity.” This indicates a strong interest in public health, disease prevalence, and their relationships with other risk factors in the population.

Green cluster (upper right): this cluster is related to experimental and pharmacological research. The terms “nonhuman,” “animal experiment,” “rats,” “rabbits,” “double blind procedure,” “drug tolerability,” and “drug efficacy” are prominent. This suggests a line of research that uses animal models and controlled clinical trials to test the efficacy and safety of new treatments.

Yellow cluster (center): this is the core of the research. Although not as large, it contains the most central terms such as “human,” “controlled study,” “clinical trial,” and “statistical analysis.” This cluster serves as the connector among the others, confirming that cardiology research in Cuba is primarily based on clinical and controlled human studies.

In summary, the graphs reveal that cardiology research in Cuba is multifaceted, with a strong emphasis on clinical research into cardiovascular disease and hypertension. Topics such as epidemiology, electrophysiology, and treatment evaluation are addressed, along with an experimental research line in animal models.

Recap

- Correlation graphs are visual representations of relationships between variables or bibliometric elements (e.g., authors, keywords, institutions, journals, etc.).
- They are based on network theory and enable the analysis of the structure and dynamics of scientific knowledge.
- A graph is composed of nodes (entities) and edges (links) that reflect relationships or associations between these elements.
- In bibliometrics, nodes can represent authors, articles, keywords, or countries, and edges can represent their degree of correlation, co-occurrence, or co-citation.
- Graph analysis allows us to identify patterns of collaboration, thematic affinity, and cognitive structures in scientific production.
- The strength of correlation between two nodes is quantified using statistical measures such as Pearson’s or Spearman’s correlation coefficient.
- There are different types of networks:
    - Co-authorship networks (collaboration between researchers).
    - Co-citation networks (articles cited together).
    - Co-word or co-occurrence networks (terms that appear together).

- Graphs can be constructed from correlation or similarity matrices derived from bibliographic data.
- The density of a network indicates the overall degree of connection between nodes.
- Centrality (degree, betweenness, closeness) measures the relative importance of a node within the network.
- Clusters or communities are groups of highly interconnected nodes that usually correspond to scientific topics or subfields.
- Modularity analysis allows these communities to be identified using clustering algorithms.
- The most commonly used tools for generating correlation graphs are VOSviewer,

Gephi, CiteSpace, BibExcel, and Pajek.

- VOSviewer constructs similarity maps and calculates distances between elements based on their degree of co-occurrence.
- Gephi offers advanced interactive visualization and structural analysis functions for large networks.
- The color and size of the nodes usually represent the intensity of correlation and the weight of the connections.
- Graphs allow us to observe the temporal evolution of topics and the emergence of new research areas.
- Proper interpretation requires combining quantitative (statistical) and qualitative (semantic) analysis of the results.
- Correlation graphs contribute to scientific monitoring, the identification of opinion leaders, and the detection of thematic gaps.
- Their correct application requires methodological rigor, careful data selection, and ethical and comprehensible visualizations.

Self-assessment questions

1. What do the nodes and edges in a correlation graph represent?
2. What type of information can be analyzed using graphs in bibliometric studies?
3. What is the difference between a co-authorship network and a co-citation network?
4. What does the correlation coefficient measure in the construction of a graph?
5. What does the density of a bibliometric network indicate?
6. What types of centrality exist, and what does each one represent?
7. What does the existence of a cluster in a correlation graph mean?
8. What tools allow correlation graphs to be constructed and scientific networks to be visualized?
9. What information is conveyed by the color and size of the nodes in a graph?
10. Why is it essential to combine quantitative and qualitative analysis when interpreting bibliometric networks?

BIBLIOGRAPHY

1. Newman MEJ. Networks: An introduction. Oxford: Oxford University Press; 2010. ISBN: 9780199206650.

2. Barabási AL. Network science. Cambridge: Cambridge University Press; 2016. ISBN: 9781107076266.

3. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. In: Cronin B, Sugimoto CR, editors. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. Cambridge (MA): MIT Press; 2014. p. 197–228. doi: 10.7551/mitpress/9780262026792.003.0010.

4. Chen C. Mapping scientific frontiers: The quest for knowledge visualization. 2nd ed. London: Springer; 2017. doi: 10.1007/978-3-319-32043-4.

5. Leydesdorff L, Welbers K. The semantic mapping of science: Co-words, co-authors, and co-citations. J Informetrics. 2011;5(1):1–14. doi: 10.1016/j.joi.2010.10.002.

6. Hanneman RA, Riddle M. Introduction to social network methods. Riverside (CA): University of California; 2005. https://faculty.ucr.edu/~hanneman/nettext/.

Chapter 7 / Capítulo 7

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch07

Heat Maps: Representing Research Density / Mapas de Calor: Representando la Densidad de la Investigación

7.1. Interpretation of heat maps in bibliometric analysis

Heat maps are a powerful visual tool that transforms multidimensional bibliometric data into intuitive representations where color intensity encodes the density, frequency, or impact of research activity. Unlike networks that emphasize structural relationships, heat maps capture patterns of thematic, geographic, or temporal concentration through color gradients, enabling rapid identification of areas of high productivity, specialized niches, and knowledge gaps. Proper interpretation of these maps requires understanding that each shade represents a continuous value, where warm colors (reds, oranges) typically indicate high density or frequency. In contrast, cool colors (blues, greens) indicate areas of lower research intensity.

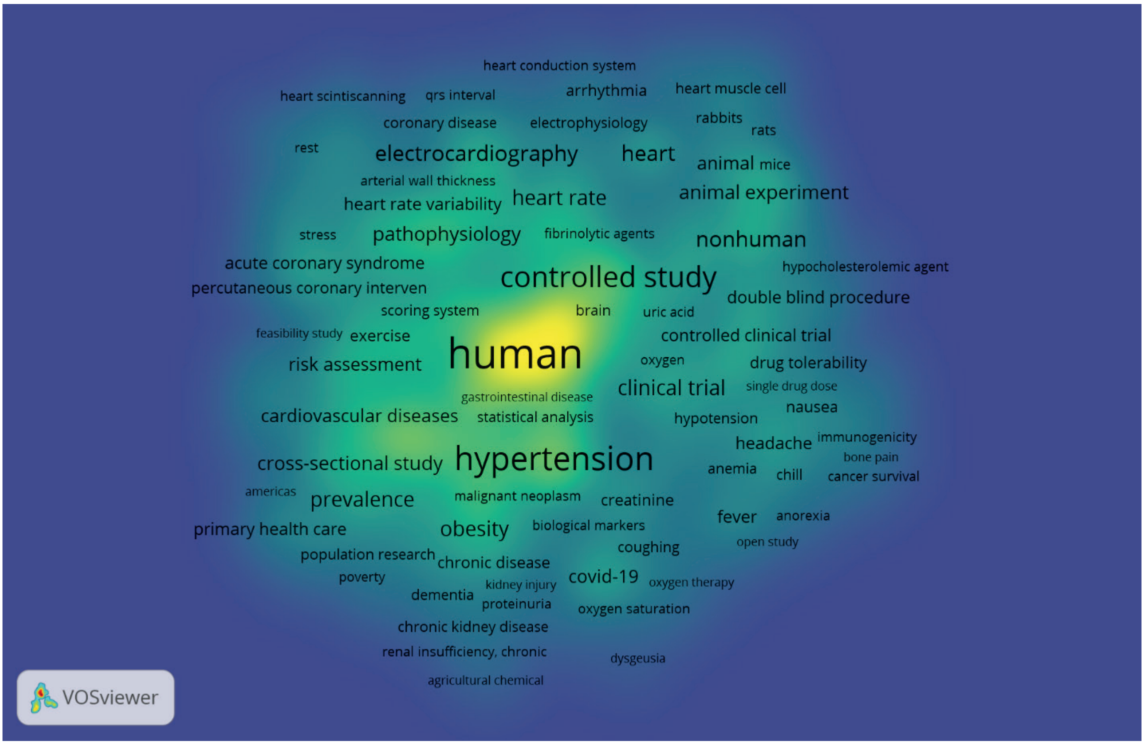


Figure 7.1. Example of a density map

Stratified reading of a bibliometric heat map involves breaking the visualization into multiple layers of meaning. The first layer, purely quantitative, reveals basic frequency distributions, where publications are concentrated, which topics receive the most attention, and which institutions lead production. The second layer, temporal, shows diachronic evolution when the map incorporates a chronological dimension, allowing us to track the migration of research interests, the emergence of new lines, and the decline of established paradigms. The third layer, relational, emerges when the heat map is superimposed with structural networks, revealing how high-density areas relate to central nodes and interdisciplinary bridges.

Contextualized interpretation transcends immediate visual analysis to integrate socio-institutional, epistemological, and science-policy factors that explain the observed patterns. An area of high research density may reflect both the intellectual vigor of a promising field and the effect of concentrated funding or passing academic fads. Cold spots may indicate unexplored frontiers with innovative potential, intellectual dead ends, or areas with methodological barriers to entry. The expert analyst distinguishes these possibilities by triangulating with other sources of evidence and specialized disciplinary knowledge, transforming the heat map from a mere distributional description into a diagnostic tool for strategic research planning.

7.2. Algorithm for constructing bibliometric heat maps

The methodologically rigorous construction of bibliometric heat maps begins with the precise definition of the dimensions to be represented, typically thematic, temporal, geographical, or institutional axes, and the selection of the intensity metric appropriate for the research objective. Metric options range from fundamental productivity indicators (e.g., number of publications) to sophisticated measures of impact (e.g., field-normalized citations) and specialization (e.g., thematic concentration indices). The choice of this metric fundamentally determines the type of patterns that the map will reveal, requiring careful alignment with the research questions that motivate the analysis.

Data processing for heat map construction involves successive stages of aggregation, normalization, and smoothing. Aggregation transforms individual publication data into summarized values for the cells of the multidimensional matrix that will constitute the map. Normalization adjusts these raw values to enable meaningful comparisons across domains with different sizes, publication practices, or citation traditions, which is essential for analyzing interdisciplinary fields. Smoothing applies interpolation algorithms to create gradual transitions between adjacent cells, improving visual readability but introducing potential artifacts that the analyst must recognize and control.

Color coding represents the stage where numerical values are transformed into interpretable visual experiences. The selection of color palettes must consider principles of visual perception, accessibility for people with color blindness, and established disciplinary conventions. Sequential palettes, with variations in brightness of a single color, are ideal for representing unidirectional magnitudes, while divergent palettes, with two contrasting colors, adequately capture deviations from a central reference point.

The choice of breakpoints between color intervals can highlight or hide critical patterns, requiring explicit methodological justification based on natural statistical distributions of the data or substantive thresholds significant to the field of study.

Interpretive validation closes the construction process, ensuring that emerging visual patterns correspond to real phenomena in the research ecosystem rather than methodological artifacts. This validation involves sensitivity testing of technical decisions (thresholds, smoothing algorithms, palettes), triangulation with other representations (networks, time series), and contrast with expert domain knowledge. The mature heat map thus transcends its initial descriptive function to become an interactive interface for analytical exploration. This dynamic tool allows hypotheses about the structure and evolution of scientific knowledge to be formulated and verified through the systematic manipulation of visualization parameters and the iterative exploration of different scales of analysis.

7.3. Creation of bibliometric heat maps

The creation of density maps in VOSviewer is integrated with graph generation, requiring only a change in the visualization in the “Items” tab to the “Density Visualization” option. This transition transforms the network representation into a heat map, where areas of intense color indicate regions of high element concentration, preserving the same spatial arrangement as in the network analysis. Density is calculated using essential functions that smooth the node point distribution, creating continuous gradients in which the color, from blue (low density) to red (high density), reveals thematic or collaborative clusters. Users can adjust the map's sensitivity using the smoothing parameter and customize the color palette to optimize readability based on the specific characteristics of the analyzed dataset.

In CitNetExplorer, density maps are generated using the "Cluster Density Visualization" function, which enables users to visualize the concentration of publications across different regions of the citation network. The process involves calculating the density of connections around each node and representing it in a heat map, with areas of greater research activity appearing in warm tones. This approach is particularly valuable for identifying periods of intense citation activity within specific research trajectories, revealing paradigmatic moments in the evolution of scientific fields.

CiteSpace implements density maps through its "Spectral Density Mapping" module, which combines spectral detection algorithms with thermal representations. The user must activate the "Show Density" option in the display control panel, which generates an additional layer of color that overlaps the conventional network map. This tool allows the identification of clusters with high internal cohesion and boundaries between different schools of thought, with the particularity that density is calculated by simultaneously considering spatial proximity and thematic similarity between elements.

In the R ecosystem, density maps are created using the *termDensity()* function for thematic analysis or *the authorDensity()* function for collaboration studies. The process first requires generating the co-occurrence or collaboration matrix and then applying the *heatmap()* function to the normalized matrix. Advanced customization includes adjusting smoothing parameters using Gaussian kernels and defining specific color palettes for different density ranges. This programming-based approach offers maximum analytical flexibility but requires technical skills in matrix manipulation and R visualization.

PyBibX in Python provides density maps via its density_analysis module, which implements kernel density estimation algorithms for bibliometric distributions. The typical workflow involves loading the bibliographic dataset, calculating multidimensional coordinates through dimensionality reduction, and then generating the heatmap with *plot_density_map()*. The library allows you to adjust the kernel bandwidth to control the level of smoothing and export maps in vector formats for high-quality publications. This implementation is compelling for analyzing large volumes of data, where efficient processing and advanced customization of visualization are required.

Recap

- Heat maps are graphical representations that show the intensity or density of a phenomenon using color gradients.
- In bibliometrics, they allow you to visualize thematic areas with varying concentrations of publications or citations.
- They are based on the distribution of frequencies or correlations within matrices of co-occurrence, co-citation, or collaboration.
- Each cell or point on the map reflects the level of research activity, measured by the number of documents, citations, or links between terms.
- Color acts as a visual variable: warmer tones (reds, oranges) indicate greater

density, and cooler tones (blues, greens) indicate less activity.

- Heat maps help detect emerging, consolidating, or declining areas within a scientific field.
- They also allow us to observe the interrelationship between topics or authors, revealing clusters of high research concentration.
- They are used in both thematic analysis (keywords) and collaboration analysis (authors, countries, institutions).
- The input data is obtained from databases such as Scopus, Web of Science, or Dimensions, and exported in a compatible format (CSV, RIS, BibTeX).
- Creating a heat map requires a numerical matrix of similarities or frequencies.
- The process includes data normalization, choice of color scale, and configuration of the density range.
- The most commonly used tools for creating bibliometric heat maps are VOSviewer, Bibliometrix (R/Biblioshiny), CiteSpace, Gephi, and Excel with statistical add-ons.
- In VOSviewer, density maps represent areas of higher element concentration (nodes) with more intense colors.
- Thematic density maps facilitate the identification of conceptual nuclei and relationships between areas of knowledge.
- Institutional or geographic maps show the spatial distribution of scientific production and international collaboration.
- These maps allow comparison of scientific performance across regions or disciplinary fields.
- The use of appropriate scales avoids visual distortions and improves the interpretation of data density.
- Heat maps should be accompanied by clear legends and complete metadata explaining the color and scale criteria.
- A rigorous interpretation combines visual reading with complementary statistical analysis (frequencies, correlations, centralities).
- When used correctly, heat maps become a powerful tool for communicating bibliometric results in a visual, intuitive, and comparative manner.

Self-assessment questions

1. What do the colors represent in a heat map applied to bibliometrics?
2. What type of quantitative information is displayed in a bibliometric heat map?
3. What is the difference between a thematic heat map and an institutional heat map?
4. What databases are typically used to generate the input data?
5. What role does standardization play in map creation?
6. What programs allow you to create heat maps of research density?
7. How does VOSviewer interpret the hottest areas on the map?
8. What precautions should be taken when choosing the color scale?
9. Why is it essential to include an explanatory legend and metadata?
10. How does the use of heat maps contribute to the visual understanding of scientific output?

BIBLIOGRAPHY

1. Chen C. Mapping scientific frontiers: The quest for knowledge visualization. 2nd ed. London: Springer; 2017. doi: 10.1007/978-3-319-32043-4.

2. Börner K. Atlas of science: Visualizing what we know. Cambridge (MA): MIT Press; 2010. ISBN: 9780262014454.

3. Börner K, Polley DE. Visual perspectives: A practical guide to making sense of data. Cambridge (MA): MIT Press; 2014. ISBN: 9780262027898.

4. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523–38. doi: 10.1007/s11192-009-0146-3.

5. Klavans R, Boyack KW. Toward a consensus map of science. J Assoc Inf Sci Technol. 2009;60(3):455–76. doi: 10.1002/asi.20991.

Chapter 8 / Capítulo 8

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch08

Sankey Diagrams: Flows And Relationships Between Concepts / Mapas De Sankey: Flujos Y Relaciones Entre Conceptos

Sankey maps are a distinctive visual tool that captures flows of varying intensity between different categories or states in a bibliometric system. Unlike traditional graphs that show adjacency or connectivity relationships, Sankey diagrams visualize quantified transfers of resources, research attention, or intellectual influence using arrows whose width is proportional to the magnitude of the flow. This representation is compelling for analyzing temporal trajectories, interdisciplinary transfers, and redistributions of research focus over time.

8.1. Interpretation

Proper interpretation of these maps requires understanding that the width of the connections encodes specific quantitative values. At the same time, the spatial arrangement reveals the structure of the system's relationships.

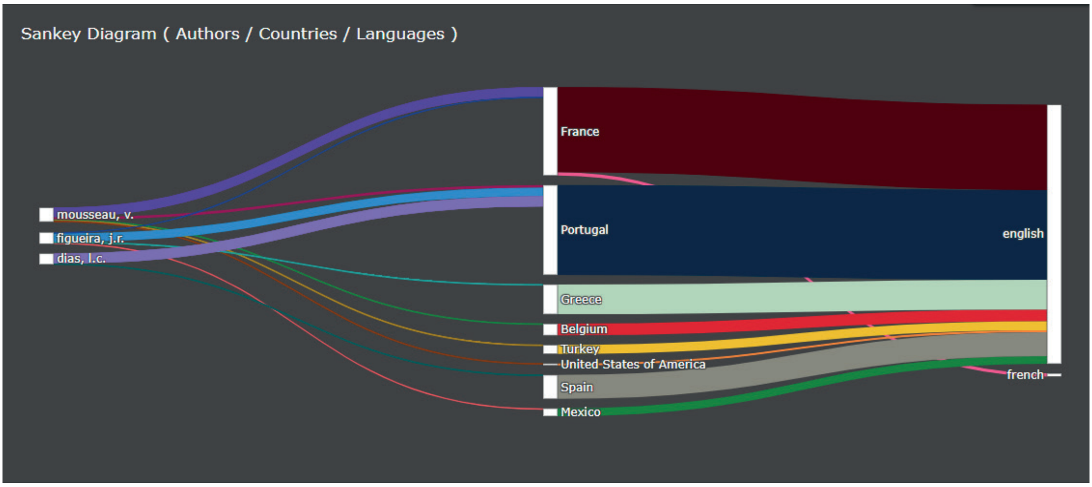


Figure 8.1. Example of a Sankey diagram

Reading a bibliometric Sankey diagram involves breaking it down into three main structural components: the nodes representing discrete categories (disciplinary fields, institutions, countries), the flows connecting these nodes indicating transfers or relationships, and the direction indicating the temporal or causal orientation of these transfers. The interpretation begins by identifying the dominant flows, those wider connections that represent the most intense relationships, to understand the main channels of intellectual exchange or institutional collaboration. Subsequently, the analysis shifts to secondary flows, which, although less intense, may reveal emerging connections or specialized niches with potential for future development. Finally, the identification of isolated nodes or those with few connections to areas disconnected from mainstream research, or to opportunities to establish new collaborations.

The temporal dimension in Sankey maps adds analytical depth by allowing the tracking of diachronic changes in flow patterns. When the diagram represents successive periods, changes in the thickness of the connections between different temporal segments reveal dynamics of consolidation, diversification, or reorientation of lines of research.

The progressive strengthening of specific flows indicates the emergence of new specialties or the intensification of established collaborations. In contrast, the thinning or disappearance of others suggests the decline of paradigms or the breakdown of research alliances. This ability to visualize trajectories makes Sankey maps an exceptional tool for analyzing scientific mobility, the evolution of research interests, and the reconfiguration of disciplinary boundaries.

Contextualized interpretation transcends immediate quantitative analysis by integrating explanatory factors that give meaning to the observed patterns. An intense flow between two disciplinary fields may reflect both fruitful integration of conceptual frameworks and mere instrumental appropriation of methodologies, without proper intellectual synthesis. The concentration of flows around specific nodes may indicate genuine intellectual leadership or scale effects derived from the size of research communities.

The expert analyst distinguishes among these possibilities by triangulating with qualitative evidence and specialized domain knowledge, transforming the Sankey diagram from a descriptive representation into a diagnostic tool for understanding the dynamics of change in the scientific landscape.

The specific bibliometric applications of Sankey maps range from the analysis of academic mobility, tracing the trajectories of researchers between institutions or countries, to the study of thematic evolution, showing how concepts migrate between different disciplinary fields. In scientific policy evaluation, they enable visualization of funding flows among agencies, thematic areas, and research teams. In interdisciplinary studies, they reveal citation patterns between distant fields or the integration of diverse methodologies.

The versatility of this representation makes it a particularly valuable tool for communicating complex findings to non-specialist audiences, translating abstract research dynamics into intuitive visual narratives about the flows that shape contemporary knowledge.

8.2. Types of Sankey diagrams

Sankey maps of thematic evolution visualize how research concepts and topics migrate across disciplinary fields or transform over time. These diagrams show the emergence, consolidation, and decline of lines of research, revealing processes of specialization, paradigm fusion, or the emergence of interdisciplinary areas. Their construction is based on the diachronic analysis of terms extracted from titles, abstracts, and keywords, segmented into successive time periods, during which the most significant terminological transitions are identified using measures of semantic similarity and evolutionary co-occurrence analysis.

Citation flow diagrams depict the transfer of intellectual influence across fields, institutions, or research groups. Unlike conventional citation networks that show specific connections, citation Sankey diagrams capture net flows of influence, showing how certain areas act as net exporters of knowledge while others function primarily as importers.

This approach is particularly valuable for studying interdisciplinary relationships, where citation flows can reveal asymmetries in intellectual exchange between established and emerging fields, or between different epistemological traditions.

Scientific collaboration Sankey maps visualize patterns of cooperation between institutions, countries, or disciplines, showing not only the existence of collaborations but also their intensity and evolution over time. These diagrams enable the identification of stable research consortia, emerging patterns of South-South or North-South cooperation, and the dynamics of collaborative network formation and dissolution.

They are typically constructed using co-authorship data segmented by period, where the flows represent the volume of joint publications between different entities, allowing analysis of how institutional collaboration strategies evolve in response to funding programs, scientific policies, or technological developments.

Funding flow diagrams depict the distribution of economic resources across different subject areas, institutions, or research types, showing how funds are allocated, transferred, and concentrated within the scientific ecosystem. These maps are handy for evaluating the impact of funding policies, identifying mismatches between stated priorities and actual allocation patterns, and analyzing the efficiency of research resource distribution. Their construction requires integrating data from calls for proposals, awarded projects, and scientific outputs, ensuring complete traceability from funding sources to the research products generated.

Sankey diagrams of scientific production by sector visualize the distribution of research outputs among different types of organizations (universities, research centers, companies, hospitals) and their evolution over time.

These diagrams capture trends in sectoral specialization in research, in knowledge transfer from academia to industry, and in the relative roles of different actors in the innovation system. Their implementation requires consistently classifying institutional affiliations by sector and analyzing their evolution over time, revealing patterns of diversification or specialization in the scientific production of different types of organizations.

Each type of bibliometric Sankey diagram answers specific research questions and requires particular strategies for data preparation, processing, and validation. The analytical objectives should guide the selection of the appropriate type, the availability of data with the necessary granularity and temporal coverage, and the study's communication needs. Regardless of the specific type, they all share the ability to transform complex bibliometric relationship data into intuitive visual narratives about the flows that structure the dynamics of contemporary scientific knowledge.

8.3. Construction

8.3.1 Implementation in R Bibliometrix

Sankey maps in Bibliometrix are constructed using the *sankeyPlot()* function, which transforms transition matrices between bibliometric categories into interactive flowcharts. The process begins with preparing a transition matrix, where the rows represent the source states, and the columns represent the destination states, with cell values indicating the frequency or intensity of the flow. For temporal thematic analysis, this matrix captures the migration of concepts across consecutive periods, while in collaboration studies, it can represent institutional mobility or changes in co-authorship patterns.

Technical implementation first requires processing bibliographic data using *biblioAnalysis()* to obtain the basic structure, then using *coupling()* or *cocMatrix()* to generate specific association matrices. The *sankeyPlot()* function accepts customization parameters such as *node. Width* to adjust the width of the nodes, *node.pad* for the spacing between elements, and *units* to specify the metric represented in the flows. Bibliometrix integrates the *networkD3* library to render interactive visualizations that allow users to drag nodes, *hover* over flows to see exact values, and filter connections by intensity thresholds.

For advanced diachronic analysis, Bibliometrix allows you to segment data into time periods using *timeslice* and generate successive transition matrices that show the evolution of flows. The *Evolution ()* theme function complements this analysis by identifying specific thematic trajectories, which can then be visualized as specialized Sankey maps. The export of results includes HTML formats to preserve interactivity, as well as SVG vector images for academic publications, maintaining visual clarity at different scales and resolutions.

The code to generate a Sankey map is:

1. Install and load packages

```
install.packages("bibliometrix")
install.packages("plotly")
install.packages("networkD3")
library(bibliometrix)
library(plotly)
library(networkD3)
library(dplyr)
```

2. Load data

```
data(management, package = "bibliometrix")
M <- management
```

3. Create data for a Sankey chart showing evolution over time

```
years <- c(2015, 2020)  # Range of years
sankey_data <- thematicEvolution(M,
                         years = years,
                         n = 10,  # Number of terms per year
                         field = «ID")  # "ID" for Keywords Plus, "DE" for Author Keywords
```

4. Generate Sankey plot

```
sankey_plot <- plotThematicEvolution(sankey_data$Nodes, sankey_data$Edges)
sankey_plot
```

8.3.2. Implementation in PyBibX

PyBibX constructs Sankey maps using the SankeyDiagram class, which provides granular control over all aspects of the visualization. The basic implementation requires creating a diagram instance by specifying the flow data as a list of tuples or a nested dictionary, where each entry defines an individual flow with its source, destination, and magnitude. The *generate()* method processes this data and applies automatic positioning algorithms that minimize connection crossings and optimize the diagram's overall readability.

Advanced customization in PyBibX includes aesthetic adjustments using parameters such as *color_palette* to define color schemes consistent with the study domain, *node_alignment* to control the vertical arrangement of nodes, and *flow_opacity* to manage the visual overlap of multiple flows. Native integration with Plotly allows you to create interactive visualizations where users can explore specific flows, dynamically rearrange nodes, and apply real-time filters based on different magnitude or category criteria.

For specialized bibliometric analysis, PyBibX includes methods such as *author_mobility_sankey()*, which transforms institutional-affiliation data into academic mobility maps, and *concept_evolution_sankey()*, which visualizes terminological migration across research periods. The tool also offers preprocessing functions, such as *normalize_flows()*, to adjust magnitudes based on different impact metrics, and *cluster_flows()*, to group minor connections and simplify complex visualizations without losing substantive information.

Recap

- Sankey diagrams show directed flows between categories, with the width of each band proportional to the flow magnitude.
- They allow for the visualization of thematic, citation, or funding transfers.
- Each link requires source-destination data and a magnitude.
- The directionality of the flow must be evident through arrows or layout.
- Normalization helps avoid visual saturation and improve comparability.
- Logarithmic scales help when flows are heterogeneous.
- Cleaning and unifying categories are essential before analysis.
- Hierarchical Sankeys show subcategories within larger flows.
- The order of nodes affects readability; minimize crossings.
- Colors should distinguish groups and have an explanatory legend.
- Sankeys should include verifiable units and sums.
- Popular software: R (ggalluvial), Python (plotly), D3.js.
- Interactivity improves the exploration of complex flows.
- In time series, panels by period or animations are used.
- Checking for outliers and aggregation errors increases validity.
- Grouping minor categories under “others” facilitates visual clarity.
- They complement graphs and heat maps by showing dynamic routes.
- Documenting aggregation criteria ensures reproducibility.
- Numerical tables should accompany Sankey diagrams.
- Interpreting flows in their disciplinary context prevents erroneous conclusions.

Self-assessment questions

1. What does the width of a band in a Sankey diagram represent?
2. What minimum data does a Sankey diagram require?
3. How would you show temporal evolution using a Sankey diagram?
4. Why are categories normalized before constructing them?
5. What advantages does it have over a static graph?
6. When would you use a logarithmic scale?
7. What effects does an incorrect node order produce?
8. How can visual overload be reduced in complex Sankey diagrams?
9. What tools can be used?
10. Why is it vital to document aggregation criteria?

BIBLIOGRAPHY

1. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire (CT): Graphics Press; 2001. ISBN: 9780961392147.

2. Healy K. Data Visualization: A Practical Introduction. Princeton (NJ): Princeton University Press; 2018. ISBN: 9780691181622.

3. Murray S. Interactive Data Visualization for the Web: An Introduction to Designing with D3. 2nd ed. Sebastopol (CA): O’Reilly Media; 2017. ISBN: 9781491921289.

4. Knaflic CN. Storytelling with Data: A Data Visualization Guide for Business Professionals. Hoboken (NJ): Wiley; 2015. DOI: 10.1002/9781119055259.

Chapter 9 / Capítulo 9

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch09

Other Results / Otros Resultados

Analysis of complementary dimensions in bibliometrics

Beyond conventional indicators of production and impact, bibliometric analysis encompasses essential contextual dimensions that reveal structural patterns in the scientific ecosystem. This chapter examines three dimensions that are often underestimated but critical to a comprehensive understanding of research dynamics: the linguistic distribution of scientific production, the role of journals as channels of specialized communication, and the geopolitical positioning of knowledge. Each of these dimensions provides unique perspectives on how scientific knowledge is organized, communicated, and distributed across different cultural, institutional, and geographical barriers.

9.1. Language interpretation in scientific production

Analysis of language distribution in scientific literature reveals profound asymmetries in the geopolitics of knowledge.

The absolute dominance of English as *the scientific lingua franca*, which typically accounts for between 80 % and 95 % of publications indexed in international databases, reflects not only communicative standardization practices but also consolidated academic power structures. This linguistic hegemony has significant implications for the visibility, accessibility, and impact of research produced in other languages, creating a systematic bias that marginalizes valuable scientific contributions developed in peripheral linguistic contexts. The interpretation of these patterns requires historical contextualization that considers how the expansion of English as a scientific language has been linked to processes of academic globalization and the concentration of publishing resources in English-speaking countries.[1]

The multilingual distribution of scientific output varies substantially across disciplines, reflecting different intellectual traditions and target audiences. In fields such as clinical medicine or environmental sciences, where research has immediate local application, it is more common to find publications in languages other than English that communicate findings relevant to specific contexts. In contrast, in disciplines such as physics or mathematics, standardization in English is almost absolute.

These disciplinary differences have important implications for scientific evaluation policies, which must recognize the legitimacy of publications in multiple languages when they respond to the specific communication needs of each field of knowledge.

The evolution of linguistic patterns over time shows contradictory trends: on the one hand, a progressive consolidation of English as the universal scientific language; on the other, periodic resurgences of publications in local languages driven by open science policies and recognition of epistemological diversity. A diachronic analysis of these trends allows us to identify moments of change in scientific communication practices, usually linked to transformations in editorial policies, the emergence of multilingual repositories, or changes in academic evaluation systems that differentially value publications in different languages.

9.2. Interpretation of the role of scientific journals

Academic journals are fundamental institutions that structure scientific fields through their functions of validation, dissemination, and community building. Bibliometric analysis of journals transcends simplistic impact metrics to examine how they shape intellectual trajectories, enshrine theoretical frameworks, and establish disciplinary boundaries. The distribution of scientific output typically follows patterns of concentration, in which a small core of journals captures disproportionate attention, while an extensive periphery hosts specialized contributions with segmented audiences.

The invisible architecture of scientific fields is revealed by mapping journal networks through co-citation and bibliographic coupling analyses. These analyses identify clusters of publications that represent distinct epistemological subfields or methodological traditions, showing how specialized communication is organized within each discipline.

The relative position of a journal within these networks provides more nuanced insights into its intellectual influence than any one-dimensional metric, revealing its role as a bridge between communities, a bastion of orthodoxies, or a space for disruptive innovation.

The ecology of journals is undergoing profound transformations driven by digitization, open science, and new economies of academic attention.

The emergence of interdisciplinary mega-journals challenges established disciplinary taxonomies, while thematic fragmentation creates spaces for ultra-specialized publications. Simultaneously, the transition to open-access models is reconfiguring power relations among publishers, authors, institutions, and scientific societies, with as yet uncertain consequences for the quality, diversity, and sustainability of academic communication.

Health indicators for the journal ecosystem must consider dimensions beyond citation impact, including geographic and linguistic diversity, equity in editorial processes, and economic sustainability. Advanced bibliometric analysis allows for the diagnosis of distortions such as citation endogamy, thematic homogenization, and excessive concentration, informing policies for a more robust, diverse, and responsive publication ecosystem that meets the evolving needs of scientific communities.

9.3. Interpretation of distribution by country

The geographical distribution of scientific output reveals profound inequalities in research capacity at the global level, with a small group of countries concentrating the majority of high-impact publications and citations. These patterns of concentration reflect historical asymmetries in the distribution of research resources, the capitalization of scientific infrastructure, and the capacity to train advanced human capital. However, diachronic analysis shows gradual convergence processes driven by aggressive science policies in emerging economies and the growing internationalization of research collaboration.

Patterns of thematic specialization by country constitute a particularly revealing analytical dimension of comparative advantages in different national innovation systems. Some countries develop publication profiles that are highly specialized in specific fields where they have competitive advantages based on natural resources, established intellectual traditions, or specific industrial clusters. Others show more diversified profiles that reflect deliberate strategies for developing general scientific capabilities. Analysis of these specialized niches enables identification of excellence niches and opportunities for complementary international collaborations.

International collaboration networks represent the most dynamic dimension of contemporary knowledge geopolitics. Bibliometric analysis of international co-authorship reveals complex patterns of preferential association that typically reflect linguistic, historical, or geographical proximities, but also the emergence of new strategic alliances based on thematic complementarities or shared resources.

The position of countries within these global collaboration networks, as central hubs, peripheral bridges, or isolated actors, significantly conditions their ability to access frontiers of knowledge and participate in transformative innovations.

9.4. Interpretation of gender in scientific authorship

A gender analysis of scientific authorship reveals deep patterns of differential participation in knowledge production. Gender distributions vary significantly across disciplines, reflecting particular histories of inclusion/exclusion, specific disciplinary cultures, and different paths to professionalization. In fields such as nursing or education, female authorship tends to be in the majority, while in engineering or physics, marked underrepresentation persists.

These disciplinary differences must be interpreted in light of historical, social, and institutional factors that have shaped the access and permanence of different genders within each field.

The authorship position is a crucial analytical dimension, as it reflects hierarchies of intellectual contribution and leadership in research projects. Consistent studies show that women are overrepresented in intermediate authorship positions and underrepresented as first authors or corresponding authors, particularly in high-impact international collaborations. These disparities point to possible barriers in the recognition of scientific leadership and in the assignment of central roles in research projects. These aspects deserve attention in gender equity policies in science.

The evolution of participation by gender over time shows complex trajectories with notable progress but also persistent stagnation. While some disciplines have experienced significant convergence, others continue to exhibit persistent gaps, particularly at the highest levels of productivity and impact. Diachronic analysis allows us to identify turning points linked to equity policies, cultural changes, or institutional interventions, providing valuable evidence for designing effective strategies to promote gender equality in the scientific system.

9.5. Interpretation of the documentary typology

The typological diversity of scientific publications reflects the plurality of discursive genres and communication forms within each discipline. Empirical research articles dominate in the natural and medical sciences, while books and book chapters remain central in the humanities and social sciences. These typological differences respond to epistemological traditions, knowledge validation practices, and academic reward structures specific to each field, and must be critically considered in any evaluative exercise.

Analysis of typological distribution over time reveals transformations in scientific communication practices.

The growing predominance of research articles in many disciplines reflects homogenization processes driven by evaluation systems based on journal metrics. However, alternative forms such as preprints, datasets, software, and other research products are resurging in response to open science movements and in recognition of the diversity of contributions to advancing knowledge.

The relative impact of different document types varies substantially across disciplines and subfields. While systematic review articles receive greater citation attention in some areas, methodological or theoretical contributions are more influential in others. Understanding these variations is essential for fair bibliometric evaluations that recognize the differential value of different types of scientific contributions according to the validation criteria specific to each epistemic community.

9.6. Interpretation of keyword analysis

Keyword analysis provides a privileged window into the conceptual and cognitive structure of scientific fields. The frequencies of terms reveal the central concepts that define a research domain, while co-occurrences show the semantic connections that structure the conceptual space. This analysis allows us to identify consolidated thematic nuclei, interfaces between subfields, and conceptual gaps that represent opportunities for innovative research.

Terminological evolution captures processes of conceptual change, the emergence of new paradigms, and transformations in the theoretical frameworks of disciplines. Diachronic analysis of keywords enables us to track the migration of concepts across fields, the emergence of specialized vocabularies, and the disappearance of obsolete terms. These patterns of terminological change reflect deeper dynamics of reconfiguration of the cognitive landscapes of disciplines.

The analysis of terminological diversity offers insights into the degree of specialization or interdisciplinarity of a field of research. Highly specialized fields exhibit dense, shared technical vocabularies, while interdisciplinary areas exhibit greater terminological heterogeneity and conceptual borrowing from multiple domains. This analytical dimension is particularly valuable for studying processes of disciplinary fragmentation, the emergence of hybrid fields, and dynamics of scientific convergence.

9.7. Analysis of institutional collaboration networks

Institutional collaboration networks reveal the organizational architecture of the scientific system, showing how different types of institutions (universities, research centers, companies, hospitals) connect to produce knowledge. The analysis of these networks allows us to identify leading institutions that function as centers of collaboration, patterns of preferential association based on geographical proximity or thematic complementarity, and the emergence of strategic consortia that reconfigure the research landscape.

The evolution of collaboration networks shows trends toward increasing institutional interconnection, but with patterns marked by structural inequalities. Institutions in central countries typically occupy more central positions in global networks, while peripheral institutions face greater barriers to integrating into international collaborations. Diachronic analysis of these networks allows us to evaluate the impact of scientific policies designed to promote more inclusive and diverse collaborations.

Institutional diversity in collaborations is an indicator of the health of the research ecosystem. Collaborations involving multiple types of institutions (academia, industry, and government) typically generate knowledge with greater potential for social and economic impact. The analysis of this collaborative diversity informs policies aimed at strengthening links between different sectors of the innovation system.

9.8. Funding patterns in scientific production

Analysis of funding agencies and programs reveals the economic infrastructure that supports scientific research. The distribution of funding sources shows the centrality of certain agencies in supporting research, the thematic priorities that guide R&D investments, and the concentration or diversification of sources of support for science. This analysis provides valuable evidence for evaluating the effectiveness of scientific funding policies.

The differential impact of different funding types is an analytical dimension of growing interest. Bibliometric studies show that competitively supported research typically achieves greater visibility and impact than research without explicit funding. In addition, the amount, duration, and type of funding (individual vs. collaborative, basic vs. applied) correlate with distinct patterns of productivity and impact, providing crucial information for optimizing research investment strategies.

Transparency in funding sources has increased significantly in recent decades, facilitating the analysis of the relationships between R&D investment and scientific results. This transparency allows connections to be drawn between funding priorities and emerging research topics, contributing to more robust accountability of science support systems and better alignment between scientific investment and social needs.

9.9. Combinations

The true power of contemporary bibliometric analysis emerges when multiple complementary dimensions are integrated to construct multidimensional analytical narratives. These combinations allow us to transcend the limitations of one-dimensional analyses, revealing complex interactions between different aspects of the scientific ecosystem. The systematic integration of dimensions such as gender, document type, collaboration patterns, and funding generates perspectives that would be impossible to obtain by examining each dimension separately, facilitating a more holistic and nuanced understanding of research dynamics.

The combination of gender analysis with institutional collaboration studies reveals crucial patterns of differential participation in scientific networks. For example, it is possible to examine whether female researchers participate in collaborative networks with the same structural characteristics as their male colleagues, or whether there are systematic differences in the type of institutions with which they collaborate. This integration enables identifying whether specific collaborative configurations, such as international multi-institutional consortia, exhibit higher or lower levels of gender equity, which is valuable for designing more inclusive scientific collaboration policies.

The intersection between typological analysis and temporal evolution reveals profound transformations in scientific communication practices. By combining these dimensions, we can track not only changes in production volumes but also transformations in the predominant discursive genres in different periods. This integration allows us to identify moments of paradigmatic change in which new forms of scientific communication emerge, or processes of homogenization in which certain types of documents displace others, with implications for the epistemological diversity of disciplines.

The triangulation of keyword analysis with funding data reveals the connections between funding priorities and the evolution of research agendas.

This combination allows us to examine how the introduction of new funding programs is reflected in the emergence of specialized terms in the literature, or how changes in funding agencies' priorities reorient the thematic focus of entire scientific communities. This analysis is particularly valuable for assessing the impact of science policies on the configuration of disciplinary cognitive landscapes.

The integration of geographical, thematic, and collaborative dimensions generates complex maps of the geopolitics of knowledge. By combining these dimensions, it is possible to identify not only which countries produce knowledge in which areas, but also how global collaboration networks are structured around different thematic specialties.

This integration reveals patterns of cognitive dependency, regional excellence niches, and asymmetries in countries' capacities to influence global research agendas.

Multivariate analysis of complementary dimensions faces significant methodological challenges, particularly in handling complex interactions and identifying causal relationships. However, approaches such as multiple correspondence analysis, structural equation modeling, and machine learning techniques allow these interrelationships to be systematically explored. The development of interactive panels that will enable different dimensions to be dynamically crossed represents a promising innovation for the visual exploration of these complexities.

The interpretation of integrated analyses requires particular sensitivity to disciplinary and contextual specificities. Patterns that appear universal may fade when examining individual disciplines, while relationships that seem marginal at the aggregate level may prove crucial in specific fields.

This contextual sensitivity is essential to avoid undue generalizations and to produce knowledge that respects the epistemological diversity of the scientific universe.

Combinations of dimensions open up possibilities for new synthetic metrics that capture multidimensional aspects of scientific production. Indicators of cognitive diversity, equity in collaborations, or alignment between funding and results emerge naturally from these integrations, offering more sophisticated tools for scientific evaluation.

Obtaining data for multidimensional bibliometric analysis is a highly contextual process that varies substantially depending on the specific research objectives and analytical questions posed. Each researcher must design a customized collection strategy that responds to their particular needs, typically starting from two primary sources: CSV files exported directly from bibliographic databases such as Scopus, Web of Science, or Dimensions, or raw results generated by specialized tools such as Publish or Perish, Bibliometrix, or VOSviewer. This initial choice profoundly shapes subsequent analytical processing, as each source has distinct formats, structures, and levels of complexity.

Once all the results are ready and you understand what each one means, the next step is to communicate them. The following section will address how to write a bibliometric report and prepare its presentation.

9.10. Visualization strategies for different bibliometric dimensions

The selection of appropriate visual representations transcends the merely aesthetic and becomes a fundamental methodological decision that directly affects the understanding and interpretation of bibliometric findings. Each type of analysis requires specific formats that highlight its characteristic patterns, facilitating the extraction of significant elements from complex datasets. The communicative effectiveness of a bibliometric study depends critically on this strategic choice, as the correspondence between the data structure and the visual format determines the recipient's ability to decode the information presented.

For analyzing linguistic distribution in scientific production, vertical or horizontal bar charts offer an optimal solution, enabling immediate comparisons of publication frequencies by language. This representation clearly visualizes the predominance of English as *a* scientific *lingua franca* over other languages, while allowing identification of temporal trends through historical series. The incorporation of color coding by language families or geographic regions adds analytical layers, transforming a simple quantitative distribution into a tool for exploring cultural hegemonies in global scientific communication.

The representation of geographical distribution finds its most intuitive and effective expression in choropleth maps, transforming tabular data into immediately understandable spatial narratives. This cartographic modality assigns color intensities according to the density of scientific production by country or region, revealing patterns of knowledge concentration and North-South asymmetries with immediate visual impact. Overlaying indicators of international collaboration through directed flows creates multidimensional visualizations that simultaneously capture individual productivity and cooperation networks, offering an integrated perspective of the global scientific landscape.

Specialized bibliometric tools such as Bibliometrix in R and PyBibX in Python automatically generate advanced visualizations that integrate multiple dimensions addressed in this chapter, combining them with bibliometric indicators such as the gross or annual citation rate.

The choice of which type of tool and representation to use is determined by the needs of what you want to communicate, always bearing in mind that you should use the most straightforward and most intuitive form of representation possible (plain text -> table -> graphic representation), without repeating information.

Recap

- Contextual dimensions complement the understanding of bibliometric results.
- The language of publication influences international visibility.
- Document type determines representativeness and weight.
- It is essential to standardize institutional affiliations.
- Gender analysis requires ethical and transparent methods.
- Gender inferences must be substantiated and documented.
- Country-specific analyses allow for the identification of regional inequalities.
- Linking funding to production helps evaluate impact.
- Industry-academia collaboration is measured in co-authorships and patents.
- Analysis by the journal shows editorial concentration.
- Crossing dimensions reveals structural inequalities.
- Missing data limits accuracy.
- Complementary sources (Crossref, OpenAlex) improve metadata.
- Geographic visualizations facilitate global comparisons.
- Analyzing languages in titles/abstracts measures accessibility.
- Database coverage influences biases.
- Normalization by field and year is essential for comparison.
- Uncertainty and data gaps must be reported.
- The confidentiality of subgroups must be respected.
- Incorporating additional variables enriches interpretation.

Self-assessment questions

1. Why is it useful to study the language of publication?

2. What biases arise from the exclusive use of Scopus or WoS?
3. How are affiliations normalized?
4. What ethical precautions are there in inferring gender?
5. How would you link funding data with topics?
6. What are the implications of document typology?
7. How can errors due to missing data be avoided?
8. Why use complementary sources?
9. What does crossing dimensions (country, gender) contribute?
10. Which visualizations are most effective?

BIBLIOGRAPHY

1. Gingras Y. Bibliometrics and Research Evaluation: Uses and Abuses. Cambridge (MA): MIT Press; 2016. ISBN: 9780262337663.

2. De Bellis N. Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics. Lanham (MD): Scarecrow Press; 2009. ISBN: 9780810867130.

3. Rousseau R, Egghe L, Guns R. Becoming Metric-Wise: A Bibliometric Guide for Researchers. Cambridge (MA): Chandos/Elsevier; 2018. ISBN: 9780081024744.

4. Waltman L. A Review of the Literature on Citation Impact Indicators. J Informetrics. 2016;10(2):365–91. DOI: 10.1016/j.joi.2016.02.007.

BIBLIOGRAPHIC REFERENCES

1. UNESCO. The race against time for smarter development | 2021 Science Report. https://www.unesco.org/reports/science/2021/en

Part IV / Parte IV

PUBLISHING AND COMMUNICATION

PUBLICAR Y COMUNICAR

AG EDITOR

Chapter 10 / Capítulo 10

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch10

Academic Writing / Redacción Académica

The culmination of any bibliometric research lies in its ability to communicate findings clearly, rigorously, and persuasively. This chapter focuses on the practical aspects of academic writing, structuring articles, and effectively communicating bibliometric results to different audiences. From the initial conceptualization of the study to the final presentation of results, each chapter provides concrete tools for transforming data analysis into meaningful scientific contributions that enrich our understanding of the dynamics of knowledge.

10.1. The idea of bibliometrics

A bibliometric study differs from other forms of research in its quantitative approach to analyzing the production, dissemination, and impact of scientific knowledge through specific metrics and statistical techniques. The conception of bibliometric research, like all scientific research, begins with the identification of a gap in the understanding of the structure or dynamics of a scientific field, where the analysis of patterns in the literature can offer solutions. Unlike systematic reviews that synthesize substantive findings or qualitative studies that explore meanings and experiences, bibliometrics focuses on macroscopic patterns emerging from the body of scientific publications, using quantitative indicators to reveal underlying trends, relationships, and structures.

Clearly defining the object of study is the first crucial step in determining a bibliometric project. This involves precisely specifying the documentary corpus to be analyzed, the time periods considered, the data sources used, and the limitations inherent in these methodological decisions. A well-defined bibliometric study is characterized by research questions that cannot be answered by simply reading the literature, but instead require the systematic analysis of large volumes of bibliographic data to reveal patterns that transcend individual subjective experience. For example, while a researcher may have intuitions about trends in their field, bibliometrics allows them to quantify these trends, identify influential factors, and discover relationships that are not obvious at first glance.

The elements that distinguish a bibliometric study from other types of articles include the systematic use of quantitative indicators, the analysis of relational patterns using network techniques, the identification of temporal trajectories, and the contextualization of findings within theoretical frameworks on scientific dynamics.

A project becomes bibliometric when its main objective is to understand the architecture of knowledge, collaboration networks, citation patterns, or thematic evolution, rather than advancing substantive expertise within a particular field.

This specific methodological orientation determines both the design of the study and the structure of the resulting manuscript, requiring detailed methodological sections on data collection and processing, as well as a presentation of results that balances quantitative rigor with meaningful substantive interpretation.

Bibliometric research addresses knowledge gaps; that is, it serves a cognitive purpose, so the scientific problem will always revolve around a lack of knowledge about research trends. This is extremely important, as it should be used to set the objective and guide the entire article. All of the above is implicit in the INTRODUCTION section of the article.

10.2. Writing the methods

The METHODS section in a bibliometric study should provide a comprehensive description that allows for complete replication of the research. This methodological transparency is a fundamental pillar of scientific integrity in bibliometrics, where seemingly minor decisions during data collection and processing can significantly influence the results. The writing must balance the technical detail necessary for reproducibility with the clarity of presentation that facilitates understanding by readers from different fields. Well-documented methods not only validate the study's credibility but also advance the discipline's methodology by enabling comparison and iterative improvement of analysis techniques.

10.2.1. PRISMA diagram for transparency

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (https://www.prisma-statement.org/) systematizes the document selection process through four sequential phases: identification, screening, eligibility, and inclusion. In bibliometrics, the identification phase documents the databases consulted, the search strategies used, and the records obtained, including duplicates. Screening applies criteria based on metadata such as document type, language, or accessibility.

Eligibility evaluates full texts based on their thematic relevance and metadata quality. The final phase specifies the corpus analyzed and the exclusion criteria applied. This structure ensures transparency in the construction of the bibliographic dataset, which is fundamental to the study's validity.

The bibliometric adaptation of PRISMA incorporates specific elements such as metadata extraction and cleaning protocols. This includes processes for author normalization, affiliation standardization, and terminology unification.

It should also document algorithms for resolving ambiguities in institutional names and managing duplicate records between databases. These expanded elements are crucial to ensuring the consistency of the analytical corpus. The flow visualization should reflect these additional steps, showing how raw data is transformed into the final dataset ready for bibliometric analysis.

Technical reproducibility completes the PRISMA framework through comprehensive documentation of tools and parameters. This includes bibliometric software versions, specific analysis configurations, and processing scripts. The thresholds applied in network analysis, the clustering algorithms used, and the normalization criteria for indicators must be specified in detail. The availability of code and datasets in accessible repositories enables independent verification of results. This methodological transparency enables exact replication of the study and facilitates comparisons between similar bibliometric research.

10.2.2. Ethical protocols

Ethical aspects of bibliometrics include considerations of data scraping, the use of sensitive information, and responsibility for interpreting and communicating results. The text should explicitly state compliance with the terms of service of the databases used, the measures implemented to avoid server overload during data extraction, and the protocols followed to ensure the secure storage and responsible use of the information obtained.

When analyzing individual researcher data, procedures for preserving anonymity and avoiding potentially harmful uses should be described, especially in evaluative contexts where results could affect career trajectories.

Disclosure of conflicts of interest and funding sources completes the fundamental ethical issues that should be addressed in the methods section.

The methodology of bibliometric studies can also be supported by a flow chart of articles, specifying the number of articles found and, from there, the flow of filters or selection criteria until the final number or sample to be worked with is obtained. Each stage of the flow should record the number of documents included and excluded, along with the specific criteria applied in each filter, allowing for immediate visualization of how the initial sample is progressively reduced to form the final bibliographic corpus for analysis. This is recommended for studies that do not work with all of the articles found, i.e., those that apply sampling, which must be declared.

10.3. Writing up the results

The presentation of bibliometric results should follow a logical sequence that guides the reader from general patterns to specific analyses. For this reason, it is always advisable to begin with fundamental productivity indicators: total number of publications, temporal distribution, and annual growth rates, as well as complementary criteria such as article type, countries, etc. These fundamental data provide the essential quantitative context for interpreting subsequent analyses. These results can be presented in tables or graphs showing the temporal evolution, highlighting periods of acceleration or deceleration in research activity.

The analysis continues with the most productive authors and institutions, using tables that rank the main contributors according to productivity and impact metrics.

This includes the number of publications per author, the institutions with the highest output, and the most active countries. These tables should be presented with well-defined columns showing ranks, names, number of documents, and total citations.

Impact and citation indicators make up the third section of results, which presents citation distributions, h-indices, and normalized indicators by field. Bar charts show citations per document, while boxplots visualize citation distributions and percentile tables establish normalized comparisons. These visual elements highlight the most influential publications and predominant citation patterns, requiring special attention to scales and ranges to avoid distortions in interpretation.

Network and collaboration analysis represents a critical visual component presented through annotated graph figures.

The sequence begins with co-authorship networks that show the main collaborative clusters, followed by co-word maps that reveal the field's thematic structure. Each figure includes detailed legends explaining the meaning of colors, node sizes, and link thicknesses, maintaining sufficient resolution to ensure readability even in article-size prints.

Density and heat maps are reserved for showing spatiotemporal patterns and thematic concentrations.

These visual elements are particularly effective for communicating the diachronic evolution of themes or the geographical distribution of scientific production. The design ensures color palettes are intuitive and accessible, using gradients that progress from low to high concentrations in a logical, consistently interpretable manner.

The application of Sankey diagrams reveals thematic mobility, changes in collaboration patterns, or the evolution of research foci.

The design of these diagrams maintains sufficient space between flows to ensure readability and uses distinctive colors for different trajectories. The complexity of these graphs requires detailed explanation in captions and references in the main text.

Temporal trend analyses are presented in a separate section, where time series of emerging concepts or evolutionary indicators are shown.

Line graphs show thematic trajectories while area graphs visualize the relative prominence of different topics over time. These elements include significant reference points, such as scientific policy launches or technological milestones, that contextualize the observed changes.

The final section presents specialized results such as burst analysis, citation patterns, or interdisciplinarity studies. These complex visualizations are reserved for the end, once the fundamental patterns have been understood.

Their inclusion as complementary figures deepens specific aspects of the study, ensuring that each advanced visualization maintains a clear justification in the research objectives.

Sometimes, results are not very relevant to the objective set out at the outset, so they should be dispensed with to give priority to those that really add weight to the research. Likewise, the order may change depending on the flow of information to be followed, ensuring that it progresses from the most general and fundamental to the most specific and advanced.

The numbering of tables and figures strictly follows the order of appearance in the text, with cross-references that guide the reader smoothly between different visual elements. Each table and figure is designed for independent comprehension, with descriptive titles and self-explanatory captions. This care for visual presentation distinguishes professional bibliometric studies from mere quantitative exercises, thereby enhancing the communicative quality of the academic manuscript.

10.3.1. Selecting visualizations by audience

The choice of visualizations in bibliometric studies must be specifically tailored to the technical knowledge and interests of the target audience.

For evaluation committees and scientific managers, priority is given to basic indicator graphs such as the temporal evolution of publications and citations, simplified collaboration network diagrams, and institutional productivity heat maps. These visualizations efficiently communicate essential information on productivity, impact, and collaborations without specialized knowledge of bibliometrics, facilitating decision-making based on robust, accessible quantitative evidence.

For academic audiences specializing in the field of study, visualizations can incorporate greater technical complexity and specific terminology. Detailed co-word maps, co-citation analyses, and thematic evolution diagrams are appropriate for these contexts. Presentations for specialists may include advanced methodological parameters, specific clustering algorithms, and network centrality measures, provided that these technical elements add substantive interpretive value to the study's findings.

When the bibliometric study is aimed at editorial boards or interdisciplinary audiences, visualizations should maximize immediate clarity and communicative impact. These visualizations transform complex data into intuitive narratives of the research field's dynamics, facilitating the transfer of bibliometric knowledge into contexts of science policy and strategic planning.

10.3.2. Tables for comparisons

Tables are the optimal format for presenting systematic comparisons between bibliometric entities such as authors, institutions, journals, or countries.

The tabular design should organize information according to clear sorting criteria, typically by productivity metrics or descending impact. Each table should include columns for ranking, entity name, number of publications, total citations, and normalized indicators, such as the h-index or average impact per document, enabling immediate multidimensional comparisons.

The hierarchical structure of tables facilitates the progressive analysis of comparative results. The initial tables present general productivity rankings, while subsequent tables show breakdowns by time periods, subject specialties, or types of collaboration.

This organization allows for the identification not only of the most productive entities globally, but also patterns of specialization, differential growth trajectories, and changes in relative leadership over time. Each table should be accompanied by brief textual analyses highlighting the most relevant comparative findings.

10.4. Writing the discussions

The discussion section is the central analytical component where bibliometric findings acquire substantive meaning through contextualized interpretation. This transition from quantitative description to qualitative explanation requires the systematic integration of empirical evidence with existing theoretical frameworks on scientific dynamics.

The construction of interpretive arguments is based on methodical triangulation among one's own findings, specialized literature, and scientific theories, establishing meaningful connections between observed patterns and underlying processes within the research ecosystem. The discussion should avoid both the mere repetition of results and unfounded speculation, balancing interpretive rigor with disciplinary relevance.

The articulation between results and previous literature represents a methodological process that requires specific strategies for documentary comparison. The systematic conceptual mapping technique identifies key publications that address similar phenomena in comparable contexts, establishing critical dialogues with previous studies.

The selection of literature for comparison should include previous bibliometric research on the same subject area, qualitative studies that explore dimensions not captured by quantitative metrics, and theoretical works that provide explanatory frameworks for interpreting the patterns observed. This documentary triangulation substantially enriches the interpretive depth of the discussion.

The identification of consistencies and inconsistencies with existing literature should be carried out through structured comparative analysis.

When the results confirm previous findings, the discussion should explore the underlying mechanisms that would explain these transcontextual regularities. In light of discrepancies with prior studies, the analysis should examine methodological factors, temporal differences, or contextual particularities that could account for the observed variations. This comparative approach transforms mere coincidences or differences into opportunities to advance theoretical understanding of the bibliometric phenomena under analysis.

10.4.1. Link to scientific policies

The discussion of implications for scientific policies is an essential component that connects bibliometric analysis with practical applications in research management. The interpretation of results should explain how the observed patterns inform the design, implementation, or evaluation of interventions within the scientific system.

For example, identifying underrepresented subject areas can support recommendations for specific funding programs, while analyzing collaboration networks can inform internationalization strategies. This link between bibliometric evidence and policy action adds substantive relevance to the study.

Developing policy recommendations requires a careful balance between empirical evidence and contextual considerations. Proposals should be logically derived from the results, avoiding undue extrapolations beyond what the data allow.

The discussion should explicitly acknowledge the partial and complementary nature of bibliometric evidence within complex policy-making processes. This rigorous approach strengthens the study’s credibility and its potential to improve research systems.

The identification of plausible causal mechanisms linking policy interventions to observed bibliometric patterns significantly enriches the discussion.

For example, changes in international collaboration patterns may be related to specific scientific mobility programs, while variations in thematic productivity may be associated with established funding priorities. This theoretical elaboration transcends mere correlation to propose well-founded causal explanations that inform future science policy interventions.

Contextualizing findings within existing science policy frameworks adds analytical depth to the discussion. Contrasting them with national and international science planning documents, reports from organizations such as UNESCO or the OECD, and evaluations of specific funding programs allows the results to be placed within broader political trends.

This contextualization facilitates the identification of mismatches between stated science policy objectives and actual patterns observed in research activity.

The development of prospective scenarios based on identified bibliometric trends represents a particularly valuable discursive strategy for scientific planning. The projection of thematic trajectories, the identification of emerging opportunities for specialization, and the mapping of future scientific capabilities transform the retrospective analysis typical of bibliometrics into a tool for strategic anticipation. These prospective exercises must be clearly grounded in the data presented and acknowledge their predictive limitations.

10.4.2. Methodological limitations

The exhaustive recognition of methodological limitations constitutes an element of intellectual rigor in the writing of bibliometric discussions.

The partial coverage of databases represents a fundamental limitation that should be characterized quantitatively when possible, specifying the geographical, disciplinary, and typological biases introduced by the selection of sources.

The discussion should assess how these limitations affect the external validity of the findings and appropriately delimit the scope of the conclusions. This methodological transparency enhances the study's credibility.

Methodological decisions in data processing introduce additional limitations that deserve critical examination in the discussion. Author disambiguation algorithms, terminology normalization criteria, and thresholds in network analysis are potential sources of bias that can affect results. The discussion should explore how different methodological decisions could have altered the patterns observed, demonstrating awareness of the constructed nature of bibliometric data. This methodological reflexivity characterizes high-quality bibliometric research.

The identification of limitations should be complemented by proposals for future research directions that overcome current methodological constraints.

The discussion may suggest strategies for validating findings through triangulation with other data sources, alternative methods of analysis, or replication across different temporal or disciplinary contexts. This projection toward future methodological advances positions the study within an evolving research program rather than as an isolated exercise.

The analysis of the differential impact of methodological limitations on different types of findings adds sophistication to the critical examination.

Some limitations mainly affect quantitative productivity analyses, while others particularly distort relational structures in network analyses. This differentiated characterization allows readers to assess which findings are more methodologically robust and which should be interpreted with greater contextual caution.

10.4.3. Integration of multidisciplinary perspectives

The incorporation of theoretical perspectives from different disciplines substantially enriches the interpretation of bibliometric findings. Conceptual frameworks from the sociology of science, the economics of innovation, and science, technology, and society studies provide valuable interpretive lenses for understanding the patterns observed. This multidisciplinary integration transcends mere bibliometric description to advance toward substantive explanations of the social and institutional dynamics underlying the quantitative data.

Triangulation with qualitative evidence on the phenomena under study is a compelling strategy for enriching the bibliometric discussion.

The incorporation of interviews with researchers, analysis of science policy documents, or ethnographic studies of laboratories, when available, allows quantitative patterns to be contextualized within specific social processes. This methodological integration overcomes the inherent limitations of purely quantitative approaches in bibliometrics.

Examining the epistemological implications of bibliometric findings adds theoretical depth to the discussion. Analyzing how the observed patterns reflect or challenge established conceptions of knowledge production, the structure of scientific disciplines, or the dynamics of paradigmatic change connects bibliometrics to fundamental debates in the philosophy of science. This elevation of intellectual dialogue maximizes the theoretical contribution of the study.

10.5. Good practices in data visualization

Adequate visualization of bibliometric data requires the systematic application of graphic design principles that balance analytical accuracy with communicative clarity. Color selection should prioritize perceptually uniform palettes and ensure accessibility for people with color blindness through specific verification tools.

Color schemes should be applied consistently across all visualizations in the study, using intuitive conventions: warm colors for high values and cool colors for low values. This visual consistency facilitates comparative interpretation between different figures and graphs.

Visual hierarchy is a fundamental principle for guiding the reader's attention to the most significant elements of each visualization.

The strategic use of size, contrast, and position highlights key patterns without distorting the integrity of the underlying data. Text labels should be carefully placed to maximize readability without obscuring important information, and connectors should be used when necessary to associate text with specific graphic elements. This hierarchical organization transforms complex visualizations into intuitive visual narratives.

The choice of visualization format should be based on the nature of the data and the specific communication objectives. Bar charts are ideal for comparing magnitudes between discrete categories, while line charts effectively show temporal trends. Network visualizations capture structural relationships, and heat maps represent densities or intensities between continuous dimensions. Each formative selection should be justified by its ability to communicate the analysis's central findings efficiently.

The scalability of visualizations warrants special consideration in academic publications. Figures should be designed to maintain legibility in both print and digital formats, using sufficient resolutions to allow zooming without loss of quality.

The balance between level of detail and visual clarity is optimized through progressive simplification techniques that present general information initially, allowing exploration of details through complementary or interactive visualizations.

Contextualizing visualizations through appropriate reference elements substantially improves their interpretability. Clearly marked scales, significant reference lines, and explanatory annotations situate the data within relevant interpretive frameworks.

This contextualization should include indicators of statistical significance when applicable, as well as comparisons with disciplinary averages or reference values established in the specialized literature.

Methodological transparency in visualization requires complete documentation of all transformations applied to the raw data. Smoothing, aggregation, or normalization procedures should be explained in detail, allowing readers to understand how processing decisions affect the patterns visualized. This transparency extends to the declaration of specific software and parameters used to generate each visualization, facilitating replication and independent verification.

Iterative usability testing with target audience representatives identifies opportunities for improvement in the visualizations. This process of continuous refinement ensures that the visual elements effectively communicate the findings to readers with varying levels of familiarity with bibliometric techniques. Incorporating feedback on aspects such as color interpretation, symbol comprehension, and interactive visualization navigation optimizes the final communicative effectiveness.

Universal accessibility represents a fundamental ethical principle in bibliometric data visualization. The design must consider diverse needs by providing alternative textual descriptions, ensuring sufficient color contrast, and avoiding exclusive reliance on color to convey critical information. These accessibility considerations broaden the study's reach and impact and align bibliometric practice with contemporary standards of academic inclusivity.

10.6. Other sections in bibliometric studies

Bibliometric studies may incorporate various additional sections depending on the specific requirements of each academic journal. Conclusions are the most common component, summarizing the main findings and their relevance to the field of study. This section should offer a concise overview that goes beyond a mere repetition of results, integrating the theoretical, methodological, and practical implications identified in the discussion. Effective conclusions establish clear connections between the study's initial objectives and its achievements, and outline meaningful directions for future research.

Acknowledgments are another common section that recognizes contributions not sufficient for authorship and sources of funding. This section should clearly specify the role of each collaborator mentioned and declare any potential conflicts of interest. Acknowledging technical support, expert advice, or access to research infrastructure enhances the study's academic transparency. Funding sources are detailed using persistent identifiers when available, facilitating the traceability of institutional support.

Appendices and supplementary material complement the main narrative without interrupting its flow. This section may include detailed methodologies, processing algorithms, additional visualizations, or extensive datasets. The organization of supplementary material should follow a logical structure with explicit cross-references from the main text. Each item included justifies its presence through its value for replication or further analysis of the presented findings.

Some specialized journals require specific sections such as practical implications, graphic summaries, or data availability statements. Practical implications translate bibliometric findings into concrete recommendations for different stakeholders in the scientific system. Graphic abstracts visually synthesize the study's main contributions, facilitating its dissemination across academic and professional channels. Data availability statements detail conditions of access to the datasets used, aligning with the principles of open science.

The standardization of these additional sections varies significantly across disciplines and academic journals. Careful review of the guidelines for authors before manuscript submission ensures compliance with specific requirements. Adapting the study structure to these editorial requirements optimizes its evaluation during peer review. The next chapter will examine in depth the criteria for selecting target journals and strategies for navigating the academic publication process in bibliometrics.

Recap

- The IMRyD structure improves clarity and consistency.
- Methods should include sources, filters, and collection dates.
- Publishing data and code promotes transparency.
- Figures should be self-explanatory and reproducible.
- The abstract should consist of the objective, method, and key results.
- Supplementary tables keep the body of the text organized.
- Standardization should be explained.
- Acknowledging limitations increases credibility.
- Discussion should relate findings to previous literature.
- R Markdown and Jupyter integrate text and analysis.
- Supplementary material increases reproducibility.
- Clarity, precision, and concise style are essential.
- Avoid overinterpreting correlations.
- Follow each journal's guidelines.
- Respond to reviews professionally.
- Declare conflicts of interest.
- Include ORCID and standardized affiliations.
- Acknowledge technical contributions.
- Adapt language according to the audience.
- Organize work files systematically.

Self-assessment questions

1. What sections make up the IMRyD structure?
2. What should the Methods section contain?
3. Why publish data and code?
4. What characterizes a self-explanatory figure?
5. How should an effective abstract be structured?
6. How should limitations be described in a manuscript?
7. What advantages does RMarkdown offer?
8. How important is it to declare conflicts?
9. How can you improve the visibility of your article?
10. What should be included in supplementary material?

BIBLIOGRAPHY

1. Schimel J. Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded. Oxford: Oxford University Press; 2012. ISBN: 9780199760237.

2. Alley M. The Craft of Scientific Writing. 4th ed. New York (NY): Springer; 2018. ISBN: 9783319695437.

3. Day RA, Gastel B. How to Write and Publish a Scientific Paper. 6th ed. Santa Barbara (CA): Greenwood; 2012. ISBN: 9781619250002.

4. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chichester: Wiley; 2009. ISBN: 9780470057247.

Chapter 11 / Capítulo 11

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch11

Publish and Communicate the Article / Publicar y Comunicar el Artículo

Bibliometric research only reaches its true culmination when the cycle of scientific publication and communication is complete. This process transforms individual work into a collective contribution to the body of knowledge, allowing the academic community to access, critique, validate, and build on its findings. Formal publication is the institutionalized mechanism that guarantees the preservation, dissemination, and legitimization of the knowledge generated, transcending the temporal and geographical limitations of the original research context. Without this essential step, even the most rigorous and innovative bibliometric study remains an incomplete academic exercise, deprived of its potential to influence the field's development.

The effective communication of bibliometric results is an epistemological imperative that goes beyond mere academic requirements. It is through the publication process that findings are subjected to critical peer review, integrated into ongoing disciplinary conversations, and become the foundation for future research.

The progressive validation of a bibliometric study occurs precisely through its use as a theoretical or methodological substrate by other researchers, who replicate, extend, or question its approaches and conclusions. This dialectical process of collective knowledge construction gives scientific publication its essential character within the contemporary research ecosystem.

Selecting the appropriate publication channel is a strategic decision that significantly influences the scope, impact, and legitimacy of the bibliometric study. Different journals and communication formats reach different audiences, operate under different methodological standards, and have varying levels of prestige and influence within the bibliometric community and applied disciplines. This initial decision must be carefully aligned with the research objectives, the methodological characteristics of the study, and the substantive contributions expected to be made to the field of knowledge. The selection process involves considering multiple dimensions that will be explored in detail in the following sections.

11.1. Journal selection

Identifying potential journals for publishing a bibliometric study requires a systematic analysis of multiple interrelated criteria. The journal's thematic scope is the first essential filter, determining whether it specializes in bibliometric studies, accepts methodological research in a specific disciplinary field, or favors bibliometric applications in particular subject areas. Reviewing recent issues allows for an evaluation of the alignment between the manuscript and the current editorial profile, identifying possible changes in the editorial team's thematic or methodological preferences. This initial evaluation prevents inappropriate submissions that would otherwise be rejected immediately for not fitting the journal's scope.

Analyzing the prestige and impact of potential journals involves considering quantitative metrics alongside qualitative assessments of disciplinary influence.

Traditional bibliometric indicators such as impact factor, CiteScore, or SJR provide comparative measures of visibility and influence. Still, they must be complemented by assessments of perceived prestige within the specific bibliometric community. Indexing in relevant databases such as Web of Science, Scopus, or specialized databases is another fundamental criterion, as it determines the accessibility and discoverability of the published article. This multidimensional analysis of potential impact enables the prioritization of journals that maximize the study's visibility among target audiences.

The evaluation of editorial times and acceptance rates is a crucial practical consideration with significant implications for the timely dissemination of research. The average times from submission to editorial decision and from acceptance to publication directly affect the timeliness of findings in rapidly evolving fields. Consulting public editorial statistics, when available, or estimating them by examining the dates of recent articles provides valuable insights into each journal's operational efficiency.

This assessment must be balanced with hy considerations of quality and impact, recognizing that rigorous review processes typically require longer timelines.

Reviewing the methodological requirements and quality standards expected by each journal identifies necessary adjustments before submission. Some publications emphasize methodological innovation in bibliometrics, while others privilege substantive applications in specific disciplinary fields. Reviewing author guidelines and recently published articles reveals expectations regarding analytical depth, methodological sophistication, and contribution to the field. This proactive preparation significantly increases the likelihood of success in the peer review process by ensuring the manuscript meets the target journal's standards.

Consideration of open access policies and associated publication costs completes the feasibility analysis for journal selection. Gold, hybrid, or green access options offer distinct advantages in terms of reader reach, compliance with funder mandates, and author costs. Evaluating potential exemptions, institutional support, or inclusion in transformative agreements enables an informed decision on the economic sustainability of publication.

This practical dimension is particularly relevant in contexts with limited resources or when multiple authors have different funding capacities.

Consulting colleagues' experiences and reviews on academic platforms provides valuable qualitative insights into the editorial process of specific journals. Assessments of review quality, fairness in editorial decisions, and the professional treatment of authors complement quantitative criteria in decision-making. This collective intelligence, available through platforms such as Scopus, Web of Science, or specialized academic forums, helps anticipate potential challenges and select journals with rigorous but fair editorial processes.

The creation of a prioritized list of potential journals, typically between three and five options, is the culmination of the selection process.

This staggered strategy allows for sequential advancement toward higher-impact publications in the event of rejections, minimizing time lost between submissions. The list should reflect a realistic balance between aspirations for impact and probabilities of acceptance, considering the novelty, methodological soundness, and specific contribution of the bibliometric study. This methodical planning transforms journal selection from an arbitrary decision into a deliberate strategy to maximize the impact and reach of bibliometric research.

11.2. Adapting to the standards of the selected journal

Adapting the bibliometric manuscript to the specific standards of the target journal is a meticulous process that goes beyond mere technical format. This adaptation involves a deep understanding of the editorial expectations regarding structure, discursive style, and the methodological approach characteristic of each publication.

The process begins with a detailed study of the instructions for authors, which specify requirements regarding maximum length, section structure, reference format, and citation conventions. This preparatory phase is crucial to avoid administrative rejections and to demonstrate respect for the publication’s established standards.

Restructuring the content to align with the journal’s preferred organizational patterns represents the first substantive step in the adaptation. Some publications favor traditional IMRaD structures, while others favor more flexible formats that prioritize narrative over sectional rigidity. Analysis of articles recently published in the journal reveals these implicit structural patterns, allowing the presentation of the bibliometric study to align with the editorial team’s expectations. This structural adaptation facilitates evaluation during the review process by presenting information in formats familiar to editors and reviewers.

Adjusting the discursive style and technical level to the target audience profile ensures the manuscript’s effective communicability. Journals specializing in bibliometrics allow dense technical language and assume familiarity with advanced methodological concepts. At the same time, disciplinary publications require greater contextualization of bibliometric methods and translation of findings into substantive implications for the field. This linguistic adaptation affects everything from the choice of terminology to the level of detail in the explanation of methodological procedures, always seeking a balance between technical precision and accessibility for the intended audience.

The reformulation of tables, figures, and visual elements to meet the journal’s technical and aesthetic requirements ensures their communicative effectiveness within the specific publication context. Specifications regarding file formats, resolutions, color schemes, and labeling conventions must be meticulously implemented during this phase. Reviewing visualizations in recent issues provides additional guidance on preferred styles, acceptable levels of visual complexity, and effective text-figure integration strategies characteristic of the publication. This attention to visual detail demonstrates professionalism and facilitates final editorial production.

Adapting the reference and citation system to the journal’s specific bibliographic standards completes the formal standardization of the manuscript.

Meticulous verification of each reference according to the required style, including author format, titles, sources, and punctuation, is a minor but critical aspect in shaping the perception of the work’s quality and thoroughness.

The use of bibliographic managers facilitates this transition between styles, but requires subsequent manual verification to detect and correct automatic inconsistencies. This bibliographic precision reflects the academic rigor characteristic of high-quality research.

Adapting the abstract and keywords to the journal’s specific conventions optimizes the article’s visibility and retrieval once published. Some publications require structured abstracts with predetermined sections, while others prefer traditional narrative formats.

The selection of keywords should reflect both bibliometric technical terminology and disciplinary terms relevant to the target audience, facilitating the crossing of disciplinary boundaries when appropriate. This metadata optimization maximizes the study's potential impact by ensuring visibility to the most relevant academic communities.

Thoroughly reviewing ethical aspects and required statements completes the process of adapting to standards. Verification of compliance with specific policies on authorship, conflicts of interest, data availability, and ethical approvals, where applicable, is a fundamental requirement for editorial acceptance. The inclusion of standardized statements in the formats and locations specified by the journal demonstrates a commitment to academic integrity and facilitates regulatory compliance assessment during the editorial process. This attention to ethical detail complements the technical and substantive adaptation of the manuscript.

Implementing all adaptations requires a thorough final review to ensure internal consistency throughout the transformed manuscript. Changes in structure, style, or format can introduce inconsistencies that undermine the argument's fluidity and the exposition's clarity. This holistic review verifies that the modifications made have improved, rather than merely altered, the presentation of the bibliometric study, maintaining its intellectual integrity while optimizing its fit with the expectations and standards of the selected journal. This iterative refinement process ensures that the manuscript represents the best possible version of the research for the specific publication context.

11.3. Presentation of information

Effectively communicating bibliometric results transcends written publication and includes various modes of presentation that broaden the research's reach and impact. Each presentation format requires specific adaptation strategies that take into account the audience's characteristics, the dissemination context, and the specific communication objectives. Versatility in presentation allows bibliometric findings to be directed to multiple stakeholders, from specialized scientific committees to decision-makers with limited technical training in bibliometrics. This communicative flexibility is essential to maximize the practical utility and social impact of bibliometric research.

11.3.1. Slide presentations

Slide presentations are the most widely used format for the oral communication of bibliometric results at conferences, seminars, and institutional meetings. Effective slide design balances substantive content with visual clarity, using the principle of informational minimalism to avoid cognitive overload in the audience. Each slide should convey a central idea through strategic combinations of concise text, striking visualizations, and graphic elements that guide attention to the most relevant findings.

This visual economy facilitates simultaneous auditory comprehension during the oral presentation.

The presentation narrative should be structured in a logical progression that contextualizes, evidences, and interprets the bibliometric findings. The typical sequence begins with the motivation and research questions, continues with a succinct methodology, presents key results through progressively visualized results, and culminates with implications and conclusions. Transitions between sections should establish explicit connections that maintain argumentative coherence throughout the presentation. This narrative structure transforms complex data into an understandable and memorable research story.

Adapting the technical level to the audience's specific profile determines the presentation's communicative success.

For specialized bibliometric audiences, methodological and technical aspects can be explored in depth, whereas for disciplinary audiences, substantive implications for the field should be prioritized. The language should be modulated according to this criterion, avoiding unnecessary technical jargon or, conversely, providing essential definitions when presenting specialized bibliometric concepts. This audience sensitivity ensures that the central message transcends disciplinary barriers.

The effective integration of visual elements requires specific considerations for the projection format. Visualizations should be simplified compared to published versions, eliminating secondary details and enlarging critical aspects to ensure readability from a distance.

The strategic use of controlled animations can progressively reveal layers of complexity in multidimensional visualizations, guiding the audience's attention through the most significant patterns. This visual optimization for projection distinguishes effective slides from mere figure transfers from the written manuscript.

Preparation for interaction with the audience completes the design of successful presentations. Anticipating frequently asked questions, preparing supplementary slides with additional information, and mastering possible methodological objections are essential elements for competently handling question-and-answer sessions.

This comprehensive preparation transforms the presentation of bibliometric results from an informative monologue into a productive academic dialogue that enriches the interpretation and contextualization of the findings presented.

11.3.2. Reproducible reports with RMarkdown and Google Colab

Reproducible reports represent a contemporary standard of methodological transparency in bibliometric research, integrating code, results, and interpretive narrative into self-contained documents. RMarkdown allows the generation of dynamic reports that execute bibliometric analyses directly from R code, automatically updating results and visualizations in response to changes in data or analytical parameters. This integration ensures consistency between the reported analyses and the underlying code, facilitating verification and extension of the study by other researchers.

The structuring of reproducible reports follows modular organization principles that clearly separate the phases of data loading, preprocessing, analysis, and visualization. Each section of the report combines fragments of executable code with explanatory text that contextualizes the operations performed and interprets the results generated. This integration of code and narrative prose transforms the report from a mere static report into an executable document that comprehensively documents the complete analytical flow of the bibliometric study.

Google Colab offers a Python-based alternative for creating reproducible reports accessible through web browsers without requiring local software installation. The ability to execute code in cloud infrastructure facilitates the processing of large volumes of bibliographic data and the easy sharing of interactive notebooks with collaborators. Integration with Google Drive simplifies the management of bibliometric datasets and the versioning of analyses, which is particularly valuable in distributed, collaborative projects.

Implementing reproducibility principles in bibliometric reports includes explicit dependency management, documentation of software versions, and the use of randomization seeds for analyses involving stochastic components. These practices allow for the exact replication of analyses months or years after their initial execution, addressing fundamental concerns about the temporal sustainability of computational research.

This attention to technical detail differentiates professionally reproducible reports from mere academic scripts.

The publication of reproducible reports alongside traditional manuscripts substantially enriches the bibliometric communication ecosystem.

Platforms such as GitHub, Zenodo, or specialized institutional repositories facilitate the permanent archiving and formal citation of these complementary resources. This emerging practice sets new standards for methodological transparency in the field, allowing the bibliometric community to build more efficiently on previous work by directly accessing the underlying computational implementations.

11.3.3. Interactive dashboards with Shiny, Dash, and Streamlit

Interactive dashboards transform the visualization of bibliometric results from static representations into dynamic, exploratory experiences that allow users to customize analytical perspectives to their specific interests. Shiny (R), Dash (Python), and Streamlit (Python) are the predominant frameworks for developing interactive web applications that communicate bibliometric findings through browser-accessible interfaces. These tools enable the creation of dashboards where users can filter data, adjust parameters, and switch between visualizations without requiring programming knowledge.

Effective bibliometric dashboard design balances advanced exploratory capabilities with intuitive interfaces that minimize the learning curve for end users.

The typical organization includes control panels with filtering widgets (sliders, drop-down menus, range selectors). These main visualization areas dynamically respond to user interactions and information panels that contextualize the displayed data. This modular architecture guides users through the analytical space without overwhelming them with unnecessary complexity.

Implementing bibliometric dashboards requires specific considerations regarding the volume and structure of the underlying data.

Performance optimization through techniques such as metric pre-aggregation, strategic sampling for large-dataset visualizations, and intelligent caching ensures responsive user experiences even with extensive bibliographic corpora.

These technical considerations are critical to the success of dashboards designed to explore large-scale document collections typical in bibliometrics.

Customizing dashboards according to specific user profiles maximizes their practical utility for different stakeholders in the scientific system. Dashboards for research managers can emphasize indicators of productivity and institutional collaboration, while versions for researchers can prioritize tools for thematic exploration and identification of emerging trends. This contextual adaptation transforms generic dashboards into specialized tools that address the specific information needs of different user communities.

The deployment and maintenance of interactive dashboards are essential operational considerations for their long-term sustainability. Platforms such as shinyapps.io, Heroku, or institutional servers provide hosting infrastructure with different advantages in terms of cost, control, and scalability. Implementing strategies for automatic data updates, usage monitoring, and iterative evolution based on user feedback ensures that dashboards remain relevant and helpful beyond the initial life cycle of the research project.

11.3.4. Scientific posters for conferences

Scientific posters represent an intermediate format between written publications and oral presentations, ideal for receiving focused feedback during academic events. Effective bibliometric poster design balances information density with visual clarity, organizing content into logical columns that guide the viewer through the research narrative.

The strategic use of typographic hierarchies, high-impact visual elements, and sufficient white space differentiates professional posters from mere overwhelming compilations of text and figures.

Adapting bibliometric visualizations to the poster format requires specific modifications to ensure readability at close range. Figures should be simplified compared to the published versions, with critical elements enlarged and high-contrast color palettes optimized for printing. The inclusion of descriptive titles and self-explanatory captions allows each visualization to communicate effectively even without immediate verbal mediation by the researcher. This visual autonomy is crucial during poster sessions with high attendee traffic.

The preparation of complementary materials enriches interaction during poster sessions. Extended abstracts, contact cards with links to digital resources, and reduced versions of the poster for distribution facilitate follow-up on conversations initiated during the event. This additional layer of resources transforms the poster presentation from an isolated event into the beginning of possible future collaborations based on the bibliometric findings presented.

11.3.5. Executive summaries for decision makers

Executive summaries translate bibliometric findings into formats accessible to management and science policy audiences with limited time and specialized technical training. These documents condense complex results into concise narratives that highlight practical implications, identified opportunities, and actionable recommendations. The language should prioritize clarity over technicality, translating specialized bibliometric concepts into clear, actionable implications for research system planning and evaluation.

The structure of executive summaries follows established conventions that prioritize key messages in the initial sections, with synthetic evidence presented in subsequent sections. The inclusion of high-impact visualizations, contextualized metrics through comparisons with relevant benchmarks, and highlights of counterintuitive findings increases the communicative effectiveness of these documents. This expository economy differentiates effective executive summaries from mere superficial simplifications of complex bibliometric analyses.

The strategic distribution of executive summaries completes the process of transfer to non-academic audiences. The identification of formal and informal channels within scientific governance structures, the adaptation of formats according to specific institutional protocols, and follow-up to evaluate the use of findings maximize the potential impact of bibliometric research on decision-making processes. This practical projection transcends traditional academic communication, directly influencing the evolution of research systems.

Concluding this section on the publication and communication of bibliometric studies, the foundations are laid for research to transcend the traditional academic sphere and generate a tangible impact on society.

Careful selection of journals, adherence to editorial standards, and mastery of multiple dissemination formats are essential skills that ensure the effective transfer of bibliometric knowledge to diverse audiences and application contexts. It is precisely this ability to connect methodological rigor with real-world needs that gives way to the next exploratory dimension: the practical application of bibliometrics in emerging scenarios and its growing influence on science, technology, and innovation ecosystems.

Recap

- Choose the journal according to its scope, audience, and editorial policy.
- Read the instructions for authors before submitting the manuscript.
- Write a clear and concise cover letter addressed to the editor.
- Use preprints for early dissemination when permitted by the journal's editorial policy.
- Select suggested reviewers ethically and responsibly.
- Understand the types of peer review and their implications.
- Deposit data and code in repositories with DOIs for reproducibility.
- Monitor impact through altmetrics and downloads.
- Adapt figures and graphic materials for different audiences.
- Prepare a summary for dissemination to the press or social media.
- Properly tag the article's metadata.
- Evaluate open access costs and options.
- Keep archived versions of manuscripts and correspondence.
- Promote open science through shared data and code.
- Monitor dissemination performance on academic networks.
- Participate in forums and conferences to increase visibility.
- Ensure compliance with ethics and copyright policies.
- Avoid misleading self-promotion or metric manipulation.
- Document the post-publication communication strategy.
- Evaluate academic and social impact as part of the closing process.

Self-assessment questions

1. What criteria help you choose the right journal?
2. What elements should a cover letter include?
3. What are the advantages and risks of publishing a preprint?
4. How are suggested reviewers chosen ethically?
5. What are the benefits of depositing data and code in repositories?
6. What do altmetrics measure, and how do they differ from traditional citations?
7. What ethical considerations should be taken into account when publicly disseminating results?
8. What types of peer review exist?
9. How can visibility be increased without violating editorial policies?
10. Why is it advisable to keep all versions and editorial correspondence?

BIBLIOGRAPHY

1. Gasparyan AY, Yessirkepov M, Voronov AA, Gerasimov AN, Kostyukova EI, Kitas GD. Scientific Publishing: Process, Practices, and Ethics. Cham: Springer; 2019. ISBN: 9783030218097.

2. Thelwall M. Web Indicators for Research Evaluation: A Practical Guide. San Rafael (CA): Morgan & Claypool; 2016. DOI: 10.2200/S00733ED1V01Y201602ICR048.

3. Sugimoto CR, Larivière V. Measuring Research: What Everyone Needs to Know. Oxford: Oxford University Press; 2018. ISBN: 9780190640125.

4. Tennant JP, Ross-Hellauer T, et al., editors. The State of the Art in Scholarly Communication. Cambridge (MA): MIT Press; 2020. ISBN: 9780262539371.

Part V / Parte V

APPLIED BIBLIOMETRICS

BIBLIOMETRÍA APLICADA

Chapter 12 / Capítulo 12

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch12

Emerging Trends / Tendencias Emergentes

The contemporary bibliometric landscape is undergoing a radical transformation, driven by the convergence of disruptive technologies that are substantially expanding the frontiers of what can be analyzed and interpreted. These technological innovations not only optimize established methodologies but also fundamentally redefine the research questions that can be asked and answered through quantitative analysis of the scientific literature. From generative artificial intelligence to distributed ledger technologies, emerging trends promise to overcome historical limitations in the field while introducing new ethical and methodological challenges that the bibliometric community must address with rigor and foresight. This chapter critically examines these transformative trends, assessing both their analytical potential and their implications for the future of scientific evaluation.

12.1. Generative AI: GPT in scientific text mining

Large-scale language models such as GPT represent a disruptive innovation in bibliometric analysis by enabling the contextual understanding of scientific texts at scale. Unlike traditional text mining methods based on term frequency or co-occurrence analysis, these models capture complex semantic relationships and interpretive nuances that previously required specialized human intervention.

This capability substantially transforms bibliometric content analysis, enabling everything from the automatic classification of articles according to complex epistemological dimensions to the identification of argumentative and rhetorical patterns that transcend the mere presence or absence of specific terms. The application of these models to entire bibliographic corpora opens up unprecedented analytical possibilities for understanding the discursive evolution of scientific fields.

The implementation of generative AI in bibliometrics enables the automated extraction of analytical dimensions that are traditionally inaccessible through conventional quantitative methods. Models can automatically identify key theoretical contributions, methodological innovations, and empirical findings in scientific publications, classifying articles according to complex taxonomies without requiring predefined terminological dictionaries. This analytical flexibility is particularly valuable for studying interdisciplinary fields where specialized vocabularies evolve rapidly and where intellectual contributions transcend established disciplinary categorizations. The ability of these systems to generate analytical summaries synthesizing findings from multiple publications further amplifies their usefulness for bibliometric reviews of large volumes of literature.

The analysis of thematic trends using generative AI significantly overcomes the limitations of keyword-based methods by capturing conceptual and semantic evolutions beyond mere terminological frequency. Models can track how certain concepts acquire different meanings in different temporal or disciplinary contexts, identifying processes of theoretical recontextualization or conceptual appropriation between fields. This ability to map the semantic drift of scientific ideas provides unique insights into the intellectual dynamics of research fields, revealing how conceptual frameworks are transformed through their circulation alongside different epistemic communities.

This diachronic analysis of semantic evolution constitutes a fundamental methodological contribution of generative AI to contemporary bibliometrics.

The automatic generation of research hypotheses is another transformative application in which AI systems analyze patterns in scientific literature to identify knowledge gaps and promising research opportunities. By processing entire document corpora, these systems can detect non-obvious connections between seemingly unrelated findings, suggest innovative combinations of methodological approaches, or identify under-explored research problems with high impact potential. This capacity for creative synthesis across traditionally separate domains of knowledge and generative AI positions it as a valuable tool for strategic research planning and the identification of emerging scientific frontiers.

The evaluation of argumentative quality and methodological soundness using generative AI introduces qualitative dimensions previously inaccessible at the bibliometric scale. Models can systematically analyze the argumentative structure of publications, evaluate the adequacy of research questions and the methods used, and identify recurring methodological limitations within specific literatures.

This ability to assess quality dimensions beyond citations offers promising alternatives to scientific evaluation systems that transcend traditional impact metrics. However, it introduces significant challenges of validation and algorithmic transparency.

The implementation of generative AI in bibliometrics faces significant challenges, including methodological transparency, biases in training data, and the reproducibility of analyses. Language models exhibit a tendency to generate factual hallucinations, reproduce biases present in their training data, and exhibit inconsistent responses across different requests. These challenges require the development of rigorous validation protocols, comprehensive documentation of parameters and prompts used, and the implementation of explainability mechanisms that allow for understanding the rationale behind automatically generated classifications and analyses.

The integration of generative AI with traditional bibliometric methods establishes a hybrid analytical paradigm that combines the strengths of both approaches. Established citation and network analysis techniques provide quantitative validation for AI-generated qualitative insights, while advanced semantic analysis contextualizes and enriches the interpretation of structural patterns identified through traditional methods.

This methodological integration produces more robust and nuanced bibliometric analyses that leverage the best of both analytical paradigms, setting new standards for a comprehensive understanding of scientific dynamics.

The future development of generative AI applied to bibliometrics will likely be oriented toward models specialized in specific scientific domains, trained on comprehensive disciplinary corpora, and capable of understanding the rhetorical and epistemological conventions particular to each field. The evolution toward multimodal systems that integrate the analysis of text, images, data, and code in scientific publications will further expand the frontiers of bibliometric analysis, enabling a more comprehensive understanding of contemporary scientific communication. This progressive specialization conditions current limitations of general models while maximizing their usefulness for specific bibliometric applications.

12.2. Blockchain for decentralized metrics

Blockchain technology is emerging as a transformative solution to address fundamental challenges of transparency, trust, and decentralization in academic metrics systems. Unlike traditional centralized databases, distributed ledgers enable the creation of incorruptible metric systems in which every academic transaction—citations, reviews, contributions—is recorded in an immutable, verifiable manner by any participant in the network.

This decentralized architecture eliminates single points of failure and reduces dependence on institutional intermediaries, laying the foundation for a more resilient and reliable metrics ecosystem. The application of blockchain to bibliometrics represents a paradigm shift toward more transparent, tamper-resistant scientific evaluation systems.

The implementation of blockchain allows for the creation of decentralized records of academic contributions that capture the entire research process, not just its final products. Each contribution, from initial conceptualization and methodological design to data analysis and writing, can be recorded with an unalterable timestamp on the blockchain, creating auditable traces of authorship and intellectual contribution. This approach solves chronic problems of authorship attribution and recognition of non-traditional contributions, particularly valuable in collaborative research where multiple researchers contribute in different capacities throughout the research cycle. The granularity of contribution recording enables fairer, more accurate metrics that reflect the complex, collaborative nature of contemporary science.

Blockchain-based metrics systems introduce innovative mechanisms for verifying citations and preventing metric manipulation.

Each citation can be recorded as a transaction on the chain, creating a transparent and immutable history of intellectual influence relationships.

This distributed record enables detection of circular citation patterns, coordinated self-citation networks, and other manipulative practices that distort traditional metrics. Decentralized verification through network consensus eliminates reliance on centralized databases whose processing and cleaning algorithms often operate as black boxes inaccessible to end users.

Academic tokens and decentralized reputation systems represent another disruptive application of blockchain in bibliometrics. These systems allow researchers to accumulate reputation based on verified contributions recorded on the chain, creating portable academic capital that is independent of specific institutions. Tokenization mechanisms can align individual incentives with collective goals of the scientific system, rewarding valuable academic behaviors such as rigorous peer review, sharing research data, or replicating previous studies. This approach fundamentally transforms academic reward systems by making values traditionally implicit in scientific culture explicit and quantifiable.

Decentralized management of academic identities through blockchain solves persistent problems of disambiguation and portability of researcher profiles. Each researcher can maintain a self-sovereign identity registered on the chain, consolidating their contributions across different institutions, platforms, and time periods. This persistent, vendor-independent identity facilitates academic mobility and eliminates the fragmentation of metrics that occurs when researchers change institutional affiliations. Integration with existing ORCID systems and other academic identifiers creates a more robust, user-centric identity ecosystem.

Smart contracts enable the automation of scientific evaluation processes based on metrics verified on the chain. These self-executing protocols can implement transparent evaluation criteria and automate processes such as peer review, funding allocation, or academic promotions based on objective metrics recorded on the blockchain.

Automation via smart contracts reduces the administrative burden on scientific evaluation while increasing the consistency and transparency of decision-making. However, this automation requires careful design to avoid encoding existing biases or creating overly rigid systems that fail to capture critical qualitative dimensions of scientific excellence.

The implementation of decentralized metrics faces significant challenges of scalability, adoption, and governance. Recording every academic transaction on the chain generates massive volumes of data that must be processed efficiently. In contrast, adoption by diverse academic communities requires overcoming institutional inertia and standardizing protocols across different disciplines and regions. Governance models for these decentralized systems must balance community participation with decision-making efficiency, avoiding both capture by particular interests and paralysis by excessive deliberation.

These practical challenges require careful attention during the design and implementation of blockchain-based bibliometric systems.

The integration of blockchain with other emerging technologies, such as artificial intelligence and the Internet of Things, creates even more sophisticated and comprehensive metric ecosystems. AI systems can analyze immutable data recorded on the chain to generate more reliable insights, while IoT devices can automate the recording of research activities in laboratories and field environments.

This technological convergence enhances the capabilities of each technology while mitigating its specific limitations, creating more robust, transparent metric infrastructures aligned with actual scientific research practices across different domains and methodologies.

12.3. Image and multimodal data analysis

Multimodal image and data analysis represents an emerging frontier in bibliometrics that significantly expands the scope of what can be quantified in scientific communication. Traditionally, bibliometric studies have focused predominantly on text analysis, neglecting the rich visual content that accompanies contemporary academic publications. Images, graphs, diagrams, and visualizations are essential components of modern scientific discourse, conveying complex information that often cannot be adequately expressed through text alone. The ability to systematically analyze these visual elements opens up new dimensions for understanding how knowledge is constructed and communicated across different scientific disciplines.

Computer vision and deep learning techniques enable the extraction of meaningful information from visual elements in scientific publications at scale. Algorithms can automatically identify and classify images across different types, from statistical graphs and flowcharts to microphotographs and three-dimensional visualizations, establishing usage patterns by specific disciplines, methodologies, or time periods.

This analytical capability transforms images from mere illustrations into quantifiable data that reveal representational preferences and standards, visualization standards, and evolutions in visual communication practices within different scientific communities. The automated analysis of these visual elements complements traditional textual approaches, offering a more comprehensive perspective on visual rhetoric in science.

Multimodal analysis integrates information from different formats—text, images, tables, equations—to generate richer, more contextualized understandings of scientific content. Rather than analyzing each modality separately, multimodal approaches capture the interrelationships between different forms of knowledge representation, revealing how researchers combine diverse semiotic resources to construct persuasive scientific arguments.

This integration allows, for example, analysis of how textual descriptions relate to corresponding visualizations, or how mathematical equations are complemented by narrative explanations and graphic representations. Multimodal bibliometrics thus captures the essentially hybrid nature of contemporary scientific communication.

The detection of methodological trends through image analysis is a particularly valuable application for scientific mapping. Certain types of images serve as reliable indicators of specific methodological approaches: electron micrographs suggest certain laboratory techniques, while circuit diagrams point to particular experimental approaches in engineering. Diachronic analysis of the prevalence of different image types can reveal the adoption, consolidation, or decline of research methodologies, providing insights into the evolution of scientific practices that complement analyses based on textual terminology. This ability to track methodologies through their visual traces represents a unique contribution of multimodal analysis to bibliometrics.

The evaluation of visual quality and clarity in scientific publications emerges as another promising application of multimodal analysis. Algorithms can evaluate aspects such as image resolution, appropriateness of scale, label clarity, and the communicative effectiveness of different types of visualizations. This automated assessment of visual quality enables large-scale studies of the evolution of scientific visualization standards, the identification of best practices in visual communication, and the analysis of how visual quality correlates with scientific impact as measured by citations and other indicators.

This traditionally subjective qualitative dimension is thus transformed into an object of systematic quantitative analysis.

Multimodal data analysis faces significant methodological challenges in standardization, interpretation, and validation. The enormous diversity of visual formats in scientific disciplines complicates the development of universally applicable taxonomies and algorithms. Furthermore, interpreting visual elements often requires specialized domain knowledge, making it difficult to fully automate the analysis without losing crucial disciplinary nuances. Validating multimodal analysis results requires innovative strategies that combine automatic evaluation with human expert verification, establishing reliability standards for this new bibliometric frontier.

The integration of multimodal analysis with other emerging trends, such as generative AI and blockchain, creates compelling synergies. Multimodal language models can generate textual descriptions of visual content or create visualizations from textual descriptions, facilitating the accessibility and searchability of multimodal scientific content. Simultaneously, blockchain can provide the infrastructure to record and verify the provenance of multimodal images and data, addressing growing concerns about the manipulation of scientific images.

These technological integrations amplify the transformative potential of multimodal analysis in bibliometrics.

The future development of multimodal bibliometric analysis will evolve toward systems capable of capturing increasingly sophisticated dimensions of integrated scientific communication. The ability to analyze not only static images but also interactive visualizations, scientific videos, and immersive virtual reality representations will further expand the frontiers of what can be analyzed.

12.4. Ethics and transparency in the age of AI

The implementation of artificial intelligence systems in bibliometrics raises fundamental ethical challenges that require robust governance frameworks. The algorithmic opacity characteristic of many contemporary models creates significant risks of reproducing and amplifying historical biases in scientific evaluation, perpetuating structural inequalities. This lack of transparency undermines the legitimacy of AI-based bibliometric evaluation systems, requiring methodological approaches that prioritize the explainability and auditability of decision-making processes to maintain the academic community's trust in these emerging tools.

Ensuring methodological transparency is an essential pillar, supported by comprehensive documentation of training data, model architectures, and validation procedures. Metadata should include detailed information on the demographic and geographic composition of the datasets used, enabling identification of representational gaps that may distort results. Researchers have an ethical responsibility to disclose both the known capabilities and limitations of their systems, avoiding undue extrapolations beyond the original development and validation conditions, which facilitates critical peer review.

The protection of personal data emerges as a critical consideration when systems process sensitive information from researchers, including career trajectories and collaboration networks. Ethical frameworks must ensure regulatory compliance, implement minimization principles in data collection, and establish clear informed consent protocols.

These safeguards are critical in evaluative contexts where results can affect career opportunities and resource allocation, requiring careful balances between analytical utility and the protection of individual privacy in the research ecosystem.

Accountability for algorithmic decisions represents another central challenge, requiring transparent chains of responsibility when automated systems generate rankings or evaluate scientific merit. Effective appeal mechanisms should be established for cases where researchers or institutions consider evaluations to be unfair, complemented by ethical oversight committees that incorporate disciplinary and geographic diversity.

This distributed governance structure allows for the identification of operational risks and the establishment of protocols for responsible use that maintain institutional trust in these emerging bibliometric evaluation systems.

Proactively mitigating algorithmic biases requires systematic strategies that include developing specialized techniques, conducting regular equity audits, and diversifying development teams. Continuous assessment of the differential impact across demographic groups, institutions, and geographic regions allows for the identification and correction of emerging disparities before they become entrenched as structural injustices.

This preventive approach is essential for building bibliometric systems that reflect the values of equity and inclusion inherent in contemporary scientific enterprise in its epistemological and methodological diversity.

Ecological sustainability completes the ethical picture by accounting for the massive computational requirements that generate significant carbon footprints during model training and deployment. Developers must prioritize energy efficiency, select balanced architectures, and make environmental costs transparent, aligning technological innovation with the imperatives of environmental responsibility that characterize contemporary science. This ecological awareness represents an essential dimension of ethics applied to the development of bibliometric tools powered by artificial intelligence in the current context of the global climate crisis.

Participatory governance involves multiple actors—researchers, managers, institutional representatives—in the design and oversight of systems, ensuring that diverse perspectives inform their development. Establishing open deliberative processes on embedded values, trade-offs between objectives, and accountability standards builds social consensus around the appropriate use of these technologies. This inclusiveness mitigates the risks of technocratic imposition and ensures that systems reflect the plural values of the global scientific community in its entirety.

Training in digital ethics develops critical users capable of interpreting algorithmic results through programs that emphasize human judgment as an essential complement to automated evaluations. Cultivating informed skepticism toward counterintuitive or potentially biased results prevents uncritical applications, while strengthening community capacities to reap benefits and mitigate risks. This literacy is a necessary investment in the human capital that will use these tools, ensuring their responsible implementation in the diverse institutional and disciplinary contexts where they will be applied.

Recap

- Open science promotes transparency and reproducibility.
- The OpenAlex and Crossref platforms democratize access to metadata.
- Artificial intelligence facilitates trend analysis and prediction.
- Altmetrics extend impact measurement beyond citations.
- The use of interactive dashboards improves dynamic visualization.
- Algorithmic transparency is fundamental in AI-based bibliometrics.
- Big data enables real-time analysis of scientific output.
- Equity indicators assess diversity across gender, language, and geography.
- Integrating patents, publications, and funding improves innovation assessment.
- Reproducible tools such as R and Python facilitate replication.
- Risks: algorithmic manipulation and metric gaming.
- Manifests such as DORA and Leiden promote responsible evaluation.
- Open repositories strengthen collaborative science.
- Data interoperability (DOI, ORCID) improves traceability.
- Alternative metrics reflect social and media impact.
- Web dashboards enable continuous monitoring of indicators.
- Metric reports are evolving toward narrative visualizations.
- AI ethics becomes a priority in research.
- Predictive bibliometrics combines algorithms and trend monitoring.
- Training in analysis tools will become essential in the future.

Self-assessment questions

1. What changes does AI bring to bibliometric analysis?
2. What is the difference between altmetrics and traditional citations?
3. What open sources are relevant for new studies?
4. What are the main ethical risks of AI in scientific evaluation?

5. What does the DORA Declaration promote?
6. Why is data interoperability important?
7. What advantages do interactive panels offer?
8. What does the concept of responsible evaluation entail?
9. What tools are key to reproducible analysis?
10. How can bibliometrics predict emerging trends?

BIBLIOGRAPHY

1. Sugimoto CR, Larivière V. Measuring Research: What Everyone Needs to Know. Oxford: Oxford University Press; 2018. ISBN: 9780190640125.

2. Thelwall M. Web Indicators for Research Evaluation: A Practical Guide. San Rafael (CA): Morgan & Claypool; 2016. DOI: 10.2200/S00733ED1V01Y201602ICR048.

3. Gingras Y. Bibliometrics and Research Evaluation: Uses and Abuses. Cambridge (MA): MIT Press; 2016. ISBN: 9780262337663.

4. Wilsdon J, et al. The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management. London: HEFCE; 2015. ISBN: 9780992492124.

Chapter 13 / Capítulo 13

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch13

Academy, Government, And Industry / Academia, Gobierno E Industria

Bibliometric analysis has transcended its traditional application in academia to become a fundamental strategic tool in the three key sectors of the innovation system. This expansion responds to the growing need for quantitative evidence to inform decision-making across science policy, research and development management, and business innovation strategies. Each sector has distinct objectives, specific institutional logics, and evaluation criteria that profoundly shape how bibliometric information is used and interpreted. Understanding these sectoral differences is essential for designing appropriate evaluation systems.

The growing interconnection between these three areas through public-private partnerships, translational research projects, and evidence-based innovation policies has placed complex demands on bibliometric systems. These demands require indicators that capture not only traditional academic excellence but also social impact, technology transfer, and economic value creation. This chapter critically examines the particularities of bibliometric use across sectors, the tensions that emerge at their interfaces, and the opportunities to develop integrated approaches that respect the distinctive missions of each field.

13.1. Differences in objectives and indicators by sector

Academia has traditionally favored indicators of scientific excellence focused on the production of basic knowledge and its validation through the peer review system. The most widely used bibliometric indicators include citations, journal impact factors, and international collaboration indices, reflecting core values such as originality, methodological rigor, and contribution to disciplinary advancement. However, significant tensions persist across academic subsectors, with the humanities and social sciences exhibiting substantially different publication patterns than the natural sciences. These internal differences complicate the standardization of evaluation criteria even within the same academic sector.

The government sector uses bibliometrics primarily to inform science policy, allocate resources strategically, and evaluate the return on investment in research and development. The preferred indicators include volume of scientific production, participation in international networks, comparative thematic specialization, and contribution to national development objectives. Here, aggregate performance metrics at the institutional or national level predominate, with less emphasis on individual impact.

The fundamental tension lies in balancing indicators of international excellence with national relevance, especially in contexts where scientific internationalization may conflict with local needs.

Industry uses radically different bibliometric approaches, primarily geared toward competitive intelligence, technology watch, and intellectual property management. Typical indicators include patent analysis, mapping of technological environments, identification of key players, and detection of innovation opportunities. The time horizon is considerably shorter than in academia, with an emphasis on immediate commercial applicability and protection of competitive advantages. The characteristic tension arises between the need for business secrecy and open science practices, creating methodological challenges in capturing contributions that are not published in traditional scientific literature.

Collaborations between sectors generate specific needs for hybrid indicators that capture contributions to both knowledge advancement and technological development. These partnerships require metrics that document academia-industry co-authorship, citations of scientific literature in patents, and researcher mobility between sectors. However, significant measurement challenges remain, particularly in capturing unpublished contributions, tacit knowledge transfer, and indirect impacts on value chains.

Developing meaningful indicators for these interfaces constitutes a crucial methodological frontier for contemporary bibliometrics.

The evolution of evaluation systems reflects divergent institutional pressures in the three sectors. While academia faces criticism over impact factors, the government sector faces demands for greater accountability, and the industrial sector requires more agile tools to navigate rapidly evolving technological environments.

These pressures are generating innovations in indicators such as alternative metrics, social impact measurements, and responsible innovation indicators. The risk of mechanically adopting indicators designed for another sector underscores the importance of developing contextualized evaluation frameworks.

The harmonization of indicators across sectors represents an ongoing challenge that requires careful balances between the standardization necessary for comparability and the flexibility essential for contextual relevance. International initiatives seek to establish common languages while recognizing legitimate differences in sectoral objectives. Recent advances in the analysis of large data sets offer opportunities to develop more nuanced indicators that capture multiple dimensions of value across different application contexts. This evolution toward multidimensional approaches promises to overcome the limitations of traditional systems while respecting the diversity of missions that characterize modern innovation ecosystems.

13.2. Evidence-based national science policies

National science policies based on bibliometric evidence represent a significant advance in the strategic management of research systems. These policies use quantitative analyses of scientific output to identify national strengths, areas of opportunity, and knowledge gaps that require priority attention. Bibliometrics provides decision-makers with objective indicators on the performance of the scientific system, the country’s thematic specialization, and its relative position in the international context. This data-driven approach enables optimizing resource allocation to areas with the most significant potential for impact and aligning research investments with national development priorities. Comprehensive bibliometric diagnoses form the basis for designing effective, context-specific science policies.

These analyses examine research productivity by region, institution, and discipline, identifying clusters of excellence and emerging niches. The evaluation of international collaboration networks reveals strategic alliances and opportunities to strengthen the country’s global insertion in world science. The detection of thematic trends through diachronic analysis enables the anticipation of future development and the prospective investment in areas with growth potential. These multidimensional diagnostics provide a complete snapshot of the national scientific system.

Bibliometrics critically informs scientific planning processes by identifying installed capacities and comparative advantages. The analysis of relative thematic specialization allows resources to be concentrated in areas where the country shows distinctive strengths or can develop competitive advantages. The evaluation of scientific impact through citations and other indicators of intellectual influence guides the prioritization of funding toward research of higher quality and international relevance. These bibliometric criteria, combined with considerations of social significance, allow for the design of national research agendas that balance scientific excellence and national relevance.

The monitoring and evaluation of implemented science policies represents another crucial application of bibliometrics in public science management. Continuous bibliometric monitoring enables measuring progress toward established objectives, evaluating the return on investment in specific programs, and making evidence-based adjustments during implementation. Post-implementation impact studies reveal changes in publication patterns, collaboration, and scientific influence attributable to particular interventions.

This continuous evaluation facilitates institutional learning and iterative improvement of national science policies. International comparisons using bibliometric indicators place national performance in a global context. These comparative analyses identify countries with similar characteristics that have implemented successful policies, providing benchmarks for the design of national strategies. The evaluation of international scientific competitiveness using standardized indicators allows for the establishment of realistic goals and the monitoring of a country's relative progress. These informed comparisons avoid cognitive isolation and facilitate the learning of international best practices adapted to the specific national context.

National scientific information systems integrate multiple bibliometric sources to provide comprehensive dashboards for decision-makers. These platforms consolidate data on scientific output, patents, collaborations, and impact in accessible and visually intuitive formats. Interoperability among national and international databases enables integrated analyses that capture multiple dimensions of scientific performance and output. These visualization tools facilitate the interpretation of complex data by non-specialist audiences, democratizing access to strategic information for science policy.

Methodological challenges in applying bibliometrics to science policy include appropriately contextualizing international indicators to specific national realities. It is crucial to adapt global metrics to capture local relevance, regional impact, and contributions to national development. Complementing bibliometric indicators with qualitative data and case studies enriches the analysis and prevents quantitative reductionism.

This mixed approach produces more balanced assessments that inform comprehensive and socially responsible science policies. The evolution toward science policies based on bibliometric evidence represents a step toward more transparent, accountable, and effective management of national research systems. The growing sophistication of analytical tools enables the development of more precise, better-targeted, and evaluable policies using robust indicators. This data-driven approach helps legitimize science policy decisions among the public and the academic community, building consensus around national priorities based on objective diagnoses and rigorous analysis of the scientific landscape.

13.3. Institutional ranking and evaluation

As mentioned in Chapter 1, institutional rankings based on bibliometric indicators have become influential tools for the comparative evaluation of universities and research centers. These systems use standardized metrics that allow the performance of institutions to be placed in national and international contexts, facilitating the identification of relative strengths and areas for improvement. However, the proliferation of diverse methodologies and different emphases on evaluation criteria generates results that must be interpreted with an understanding of their methodological foundations.

Transparency in calculation procedures and clarity about the objectives of each ranking are essential for their proper use in institutional management.

The central global ranking systems, such as the SCImago Institutions Rankings, the CWTS Leiden Ranking, and the QS World University Rankings, employ distinct methodological and y approaches that reflect different conceptions of institutional excellence. While some favor indicators of productivity and scientific impact, others emphasize academic reputation, internationalization, or social impact.

These methodological differences explain the variations in the positions that the same institution may occupy in different rankings. A thorough understanding of the indicators that make up each ranking enables institutions to identify improvement strategies aligned with their particular missions and contexts.

Bibliometric-based institutional evaluation must carefully account for disciplinary differences to avoid bias against areas with distinctive publication and citation practices.

The humanities, arts, and certain social sciences exhibit bibliometric patterns that differ from those of the natural sciences and engineering, requiring methodological adjustments in evaluation systems. The normalization of indicators by field of knowledge, the use of percentiles instead of absolute values, and the consideration of diverse academic products beyond scientific articles are necessary practices for equitable institutional evaluations.

These methodological refinements enable fairer comparisons across institutions with different disciplinary profiles.

The responsible use of rankings in institutional management requires avoiding an obsession with numerical rankings and focusing on trend analysis and relative progress. Diachronic monitoring of performance on specific indicators provides more valuable information for strategic planning than mere attention to position in rankings. Identifying comparable institutions that serve as benchmarks allows for setting realistic goals and learning from successful experiences. This analytical approach transforms rankings from instruments of competition into tools for continuous improvement and institutional learning.

Complementing bibliometric indicators with other sources of information substantially enriches institutional evaluation. Integrating data on innovation, technology transfer, social impact, and student satisfaction provides a more comprehensive view of institutional performance. Multidimensional evaluation systems that balance research metrics with indicators of teaching, outreach, and management capture the complexity of contemporary university missions. This comprehensive approach prevents distortions in the allocation of resources and incentives within higher education institutions.

Well-founded criticism of ranking systems is a healthy practice that drives methodological improvements and more responsible uses.

Limitations in database coverage, geographical and linguistic biases, and excessive simplification of complex institutional realities are valid objections that must be considered when interpreting results. The development of alternative rankings with innovative approaches, such as those that prioritize sustainability, inclusion, or teaching innovation, enriches the evaluation ecosystem and responds to emerging social demands. This diversification of evaluative approaches benefits the entire higher education system.

Institutional evaluation based on bibliometric evidence, when practiced with methodological rigor and awareness of its limitations, contributes to the transparency and improvement of higher education and research systems.

The challenge lies in developing institutional cultures that leverage the perspectives these instruments offer without falling into quantitative reductionism or an obsession with ranking positions. The balance between the informed use of metrics and the preservation of institutional diversity and academic autonomy is the key to an evaluation that genuinely serves the progress of science and higher education.

13.4. University-business knowledge transfer

Knowledge transfer between universities and businesses is a fundamental process for converting scientific knowledge into innovation and economic development.

This two-way flow allows advances generated in academia to find practical applications in the productive sector, while business needs inform new directions for university research. Bibliometrics offers valuable tools for measuring, analyzing, and optimizing these transfer processes, providing indicators of the intensity, characteristics, and impact of collaborations between the two sectors. The study of these interactions using specific metrics allows for the design of more effective technology transfer policies.

Scientific publications with multiple authors from both university researchers and business professionals are a direct indicator of collaboration in research and development. Analysis of these co-authorships reveals patterns of thematic specialization, collaborative intensity, and the evolution of university-business relationships over time. Identifying the most active institutions and companies in these collaborations allows for the recognition of best practices and successful models of linkage. An examination of co-authorship networks reveals how these collaborative ecosystems are structured and which actors serve as bridges between the two sectors, facilitating the flow of knowledge.

Citations of scientific literature in patents represent another crucial bibliometric indicator for measuring the transfer of scientific knowledge to the productive sector. This analysis enables tracking how findings from basic and applied research inform the development of new technologies and products. Identifying influential scientific publications in patent law reveals the areas where university research has the most significant impact on business innovation. The study of the technological fields that make the most intensive use of recent scientific knowledge helps to prioritize areas of research with high potential for technology transfer.

Indicators of researcher mobility between universities and companies complement the analysis of collaborations, showing how human capital facilitates knowledge transfer. Tracking career paths that alternate between both sectors reveals patterns of expert circulation. The identification of researchers who maintain scientific productivity during periods in companies demonstrates the compatibility between research activity and industrial technological development. This mobility favors the creation of informal networks and the transfer of tacit knowledge, crucial dimensions of the innovation process.

Assessing the economic and social impact of transfer requires combining bibliometric indicators with other data sources.

The creation of university spin-offs, research contracts with companies, and participation in public-private consortia represent tangible results of transfer that can be correlated with bibliometric indicators. The integrated analysis of these metrics allows for the development of more comprehensive models to evaluate the return on investment in university research and its contribution to the regional and national innovation ecosystem.

Methodological challenges in measuring university-business transfer include capturing unconventional forms of collaboration and attributing causality to research and innovation outcomes. Many valuable interactions, such as consulting, specialized training, or equipment sharing, are not reflected in traditional bibliometric indicators. Complementing these with qualitative studies, researcher surveys, and specific case analyses enriches our understanding of these complex processes. This mixed methodological approach produces more balanced and beneficial evaluations for knowledge transfer management.

The evolution toward more sophisticated indicators of university-business transfer represents an important frontier for applied bibliometrics. The development of metrics that capture the bidirectionality of knowledge flow, the diversity of transfer mechanisms, and the impacts at different time scales will enable more strategic management of these collaborations. This specialization responds to the growing importance of knowledge-based innovation for business competitiveness and socioeconomic development, positioning bibliometrics as an essential tool for science, technology, and innovation policy.

13.5. Bibliometrics for social impact assessment

Assessing the social impact of research using bibliometric tools represents a significant step toward more comprehensive evaluation systems. Traditionally, bibliometrics has focused on measuring academic impact through citations, but there is a growing demand to capture how research influences society beyond the educational sphere.

This evolution responds to the need to demonstrate the social value of public investment in science and to align research activity with urgent social challenges. The development of social impact indicators is an emerging field that complements traditional metrics of scientific excellence.

Mentions in public policy documents are a valuable indicator for measuring the influence of research on government decision-making. Analyzing how scientific publications are cited in legislation, parliamentary reports, and strategic planning documents reveals the path of knowledge from academia to the public sphere. Identifying research that is particularly influential in specific policies provides insight into the characteristics that facilitate the use of scientific knowledge in public management. This analysis informs strategies for making research more accessible and relevant to decision-makers.

Presence in the media and on social networks adds another dimension to assessing the social reach of research. Alternative metrics capture how scientific findings are discussed, shared, and commented on in digital platforms and traditional media.

This analysis reveals which research topics generate the most public interest and how science is integrated into social debate. Identifying researchers and institutions that effectively communicate their work to the general public provides models for improving scientific outreach and engagement with civil society.

Collaborations with non-governmental organizations and local communities represent another facet of social impact that can be measured through adapted bibliometric indicators. Co-authorship with representatives of civil society, participatory research, and projects co-designed with communities demonstrates forms of socially rooted research. Analysis of these collaborations reveals more inclusive and socially responsible models of knowledge production. Identifying successful participatory action research practices informs strategies for building stronger bridges between academia and society.

Contributing to the Sustainable Development Goals provides an emerging framework for assessing the social impact of research. Mapping scientific publications to specific sustainable development goals allows us to visualize how research addresses global challenges such as climate change, inequality, and public health. This analysis identifies thematic gaps where more research is needed to address urgent social problems. Evaluating the contribution of research to global development agendas provides a contextualized perspective on the social value of science.

Methodological challenges in measuring social impact include causal attribution, the temporal diversity of impacts, and consideration of local contexts. The social effects of research may manifest years after publication and through indirect pathways that are difficult to trace. Complementing quantitative indicators with case studies, impact narratives, and participatory methodologies enriches the assessment. This mixed approach better captures the complexity of social impact and avoids reductionist measurement.

The evolution toward evaluation systems that equally value academic and social impact represents a necessary transformation in research culture.

The development of multidimensional frameworks that recognize different types of contributions to advancing knowledge and social well-being promotes more diverse and relevant science. This evolution responds to social demands for more responsible research that is aligned with the values and needs of the communities it serves, positioning bibliometrics as a tool for science that is more aware of its social role.

13.6. Metrics-based intellectual property management

Strategic intellectual property management has progressively incorporated bibliometric tools to optimize decisions on protection, commercialization, and technological development. These metrics enable evaluating innovation potential, identifying patenting opportunities, and prioritizing investments in research with a higher probability of generating valuable intellectual property rights. Bibliometric analysis applied to intellectual property transcends the traditional approach focused on patent counting, developing more sophisticated indicators that capture quality, impact, and competitive positioning. This quantitative approach complements traditional legal expertise in technology management.

The analysis of citations between patents and scientific publications is a fundamental metric for measuring knowledge transfer between science and technology. Citations of scientific literature in patent documents indicate how basic research feeds technological development, while citations of patents in scientific articles reveal reverse flows where technology influences research. The identification of publications that are particularly cited in patents points to research with high potential for technological application. Mapping these cross-citation networks allows us to identify fields where science and technology interact most intensely, informing innovation-oriented research strategies.

Patent value indicators based on bibliometric metrics offer insights into the commercial and technological potential of inventions. The geographic breadth of protection, measured by the number of countries in which a patent is filed, reflects perceptions of global commercial value. Citations received from subsequent patents reflect technological influence and importance for the field's further development. Provide claim density and additional indicators of value. These combined indicators enable prioritization of intellectual property portfolios and informed decisions about maintenance, licensing, or abandonment of rights.

The analysis of technological landscapes using bibliometric techniques enables mapping of competitiveness and innovation opportunities. The aggregation and analysis of large patent datasets reveal technological trends, key players, and gaps in technological development. The identification of emerging technology clusters and converging technologies informs competitive positioning strategies. Analyzing the technological diversity of specific actors allows for the identification of opportunities for collaboration or acquisition. These tools support strategic research and development decisions by providing data-driven competitive intelligence.

Evaluating the innovative performance of institutions using intellectual property metrics requires specific methodological adjustments. Normalization by institutional size, thematic specialization, and research intensity allows for fairer comparisons.

The integration of patent indicators with scientific publication metrics provides a more complete view of innovation performance. The analysis of transfer efficiency, measured through the relationship between research investment and intellectual property generation, informs the productivity of the institutional innovation system. These comprehensive assessments support the strategic management of research and innovation.

Methodological challenges in applying bibliometrics to intellectual property include differences in patenting practices across sectors and countries, as well as temporal variability in protection strategies. Pharmaceutical technologies exhibit patterns different from those of information technologies, while patenting strategies vary significantly between companies and universities. Complementing this with qualitative analysis, case studies, and specific sector expertise enriches the interpretation of metrics.

This contextualized approach avoids erroneous conclusions drawn from direct comparisons across institutional or sectoral realities.

The integration of intellectual property information systems with bibliometric platforms represents the current frontier in data-driven technology management.

The development of dashboards that combine metrics on publications, patents, collaborations, and impact provides powerful decision-making tools. The application of artificial intelligence techniques for predictive analysis of innovation potential allows opportunities to be identified before they become apparent. This evolution toward comprehensive knowledge management systems positions bibliometrics as an essential component in modern technological innovation management systems.

You have now mastered the fundamentals of bibliometrics and have the methodological tools, analytical skills, and critical perspective necessary to undertake rigorous and meaningful bibliometric research. This knowledge places you on the threshold of a vast territory to explore, where each study will represent both a practical application of what you have learned and an opportunity to contribute to the advancement of this constantly evolving discipline. The real journey begins now: take these foundations to new research questions, innovative contexts, and original contributions that expand the frontiers of quantitative analysis of science, always balancing metric rigor with the interpretive depth and ethical responsibility that characterize outstanding bibliometrists.

Recap

- The Triple Helix model describes the interaction between universities, industry, and government.
- Co-publications are indicators of scientific collaboration.
- National innovation systems measure technological efficiency.
- Technology transfer requires university liaison offices.
- Public policies guide research priorities.
- Patents reflect the practical application of knowledge.
- Entrepreneurial universities promote spin-offs and startups.
- Combined indicators (publications + patents) show innovation.
- Ethical governance prevents conflicts of interest.
- Transparency in intellectual property ensures trust.
- Tax incentives encourage public-private collaboration.
- Balanced evaluation values basic and applied research.
- Technology parks promote knowledge transfer.
- International networks strengthen global innovation.
- University-industry agreements require precise regulation.
- Investment in R&D correlates with national competitiveness.
- Socioeconomic impact is an emerging criterion in evaluation.
- Measurement systems must incorporate social value.
- Equitable partnerships with developing countries reduce gaps.
- The Triple Helix is evolving toward quadruple helix models with civil society.

Self-assessment questions

1. What does the Triple Helix model describe?
2. What combined indicators can you use to measure university-industry transfer?
3. What ethical risks exist in academia-industry collaboration?
4. What public policy instruments promote innovation (subsidies, loans, tax incentives)?
5. How can you measure the socioeconomic impact of scientific research?
6. What role do spin-offs and technology parks play in knowledge transfer?
7. Why is it important to distinguish between fundamental and applied research in evaluation systems?
8. What factors define an entrepreneurial university?
9. What criteria allow for the evaluation of equitable international collaborations with developing countries?
10. What metrics capture the triple helix (publications + patents + contracts)?

BIBLIOGRAPHY

1. Etzkowitz H. The Triple Helix: University–Industry–Government Innovation and Entrepreneurship. London: Routledge; 2008. ISBN: 9780415432305.

2. Clark BR. Creating Entrepreneurial Universities: Organizational Pathways of Transformation. Oxford: Pergamon/Elsevier; 1998. ISBN: 9780080436270.

3. Nelson RR. National Innovation Systems: A Comparative Analysis. Oxford: Oxford University Press; 1993. ISBN: 9780195076179.

4. Simon D, Kuhlmann S, Stamm J, Canzler W, editors. Handbook on Science and Public Policy. Cheltenham: Edward Elgar; 2019. ISBN: 9781788973039.

APPENDICES / APÉNDICES

I. TECHNICAL GLOSSARY

A

Altmetrics: Alternative metrics that measure the impact of research beyond traditional citations, including mentions in social media, policy, and the media.

Co-citation analysis: Technique that identifies relationships between documents based on the frequency with which they are cited together.

Co-authorship analysis: Study of patterns of collaboration between researchers through joint publications.

Self-citation: Citation by an author of their own previous work.

B

Bibliometrics: Discipline that applies statistical and mathematical methods to analyze the production and dissemination of scientific knowledge.

Citation burst: Period in which a document experiences a significant increase in its citation frequency.

C

Open science: Movement that promotes free access to publications, data, and research methods.

Scientometrics: Quantitative study of science as a social and economic system.

Citation: Reference to a previous work in a scientific publication.

CiteScore: Journal impact metric calculated by Scopus based on citations received by published documents.

Thematic cluster: Group of concepts, authors, or documents that are highly connected in a bibliometric map.

D

Disambiguation: Process of distinguishing between entities with similar names (authors, institutions).

Sankey diagram: Visualization showing flows between different groups or categories.

E

Ethical scraping: Extraction of data from web sources while respecting terms of service and technical limitations.

Scientific excellence: Percentage of an institution's publications that are among the top 10% most cited worldwide.

F

Impact factor: Metric that measures the average frequency with which articles in a journal are cited in a specific period.

FWCI (Field-Weighted Citation Impact): Indicator that compares the citations received by a set of documents with the global average in their fields.

G

Google Scholar: Free search engine that indexes academic literature from multiple sources.

H

H-index: Index that combines productivity and impact, where a researcher has an h-index if h of their articles have at least h citations each.

I

g-index: A variant of the h-index that gives more weight to highly cited publications.
m-index: h-index normalized by years of research career.
Interdisciplinarity: The degree to which a piece of research integrates methods or concepts from multiple disciplines.

J

Journal Impact Factor: See Impact factor.

L

Bradford's Law: Bibliometric law describing the uneven distribution of relevant literature across scientific journals.
Lotka's Law: Law describing the distribution of productivity among scientific authors.

M

Heat map: Visualization that uses colors to represent densities or intensities of bibliometric phenomena.
Scientific map: Visual representation of the intellectual or social structure of a field of research.
MeSH (Medical Subject Headings): Controlled vocabulary used to index articles in PubMed.
Responsible metrics: Ethical and contextualized use of bibliometric indicators.

N

Normalization: Adjustment of bibliometric indicators to allow comparisons between different fields or periods.

P

PubMed: Free bibliographic database maintained by the National Library of Medicine.
Pybliometrics: Python library for accessing and analyzing Scopus data.
Keywords: Terms that describe the essential content of a document.

R

Collaboration network: Graphical representation of cooperative relationships between researchers or institutions.
Reproducibility: Ability to replicate a bibliometric analysis using the same data and methods.

S

Scopus: Commercial bibliographic database maintained by Elsevier.
SCImago Journal Rank (SJR): Journal prestige indicator that weights citations by the prestige of the citing source.
SciELO: Electronic library that indexes scientific journals from Latin America and other countries.

T

Thesaurus: Controlled and structured vocabulary used for indexing and information retrieval.

V

VOSviewer: Software for building and visualizing bibliometric networks.
Visualization: Graphical representation of bibliometric data to facilitate its interpretation.

W

Web of Science: Commercial bibliographic database maintained by Clarivate Analytics.

II. ADDITIONAL TERMS

Academic footprint: Academic footprint representing the impact and visibility of a researcher.
Bibliographic coupling: Technique that links documents based on their common references.
Citation analysis: Analysis of citation patterns to evaluate impact and influence.
Co-word analysis: Analysis of co-occurrence of terms to identify conceptual structures.
Data mining: Extraction of patterns and knowledge from large bibliographic data sets.
Domain analysis: Study of the structure and dynamics of a specific scientific field.
Ethical bibliometrics: Application of ethical principles in bibliometric practice.
Gender gap: Gender gap in productivity, impact, or scientific recognition.
Knowledge diffusion: Process of disseminating scientific knowledge through publications and citations.
Literature mapping: Techniques for mapping the intellectual territory of a field of study.
Metric invariants: Bibliometric properties that remain constant across different contexts .
Open citations: Movement to make citation data accessible.
Predatory publishing: Publishing practices that prioritize profits over academic quality.
Quantum metrics: New generation of metrics based on complex data analysis.
Research assessment: Evaluation of the quality and impact of research using bibliometric indicators.
Science governance: Use of metrics to inform scientific policy and management.
Triangulation: Use of multiple methods or sources to validate bibliometric findings.

III. RESOURCES AND LINKS OF INTEREST

Rules and standards

Ethical Declarations and Frameworks

- DORA Declaration (San Francisco Declaration on Research Assessment): https://sfdora.org/
- Leiden Manifesto on Research Metrics: https://www.leidenmanifesto.org/
- Hong Kong Principles on Researcher Evaluation: http://www.hkprinciples.org/

Bibliometric Standards

- ISO 31-19: Science and Technology Indicators: https://www.iso.org
- NISO Alternative Assessment Metrics Initiative: https://www.niso.org/standards-committees/altmetrics
- COUNTER Code of Practice for Metrics: https://www.projectcounter.org/

Official Documentation of Tools

Bibliometric Analysis Software

- Bibliometrix (R) - Official Documentation: https://www.bibliometrix.org/
- VOSviewer - Manual and Tutorials: https://www.VOSviewer.com/documentation
- CiteSpace - User Guide: http://cluster.cis.drexel.edu/~cchen/citespace/
- CitNetExplorer - Tutorials: https://www.citnetexplorer.nl/

Programming Platforms

- R for Bibliometrics - CRAN Task View: https://cran.r-project.org/web/views/
- Python Bibliometric Libraries - PyPI: https://pypi.org/
- Google Colab - Guides: https://colab.research.google.com/

Application Downloads

Main Tools

- VOSviewer - Download: https://www.VOSviewer.com/download
- CiteSpace - Download: http://cluster.cis.drexel.edu/~cchen/citespace/download/

- CitNetExplorer - Download: https://www.citnetexplorer.nl/download.php
- Publish or Perish - Download: https://harzing.com/resources/publish-or-perish

Development Environments

- RStudio - Download: https://www.rstudio.com/products/rstudio/download/
- Anaconda (Python) - Distribution: https://www.anaconda.com/products/distribution
- Jupyter Notebook: https://jupyter.org/install

Databases and APIs

Access to Bibliographic Data

- Scopus API Documentation: https://dev.elsevier.com/
- Web of Science API: https://developer.clarivate.com/apis/wos
- Dimensions API: https://docs.dimensions.ai/dsl/
- CrossRef API: https://www.crossref.org/documentation/retrieve-metadata/
- OpenAlex API: https://docs.openalex.org/

Checklists and Guides

Research Assessment

- PRISMA Checklist for Systematic Reviews: http://www.prisma-statement.org/
- Guide to Responsible Metrics - LERU: https://www.leru.org/publications
- Research Assessment Framework - UKRI: https://www.ukri.org/

Scientific Publication

- COPE Guidelines for Publication Ethics: https://publicationethics.org/
- Guidelines for Authors - ICMJE: http://www.icmje.org/
- STROBE Checklist for Observational Studies: https://www.strobe-statement.org/

Data Repositories

- Zenodo - Research Data Repository: https://zenodo.org/
- Figshare - Open Data Platform: https://figshare.com/
- GitHub - Code and Scripts: https://github.com/

Communities and Forums

Discussion and Support

- ResearchGate - Scientific Community: https://www.researchgate.net/
- Stack Overflow - Technical Support: https://stackoverflow.com/

Professional Associations

- ISSI - International Society for Scientometrics and Informetrics: https://www.issi-society.org/
- ASIS&T - Association for Information Science and Technology: https://www.asist.org/
- LIBER - European Research Libraries Association: https://libereurope.eu/

Online Tools

Analysis and Visualization

- Litmaps - Literature Mapping: https://www.litmaps.com
- ResearchRabbit - Literature Discovery: https://www.researchrabbit.ai/
- Connected Papers - Visual Exploration: https://www.connectedpapers.com/

Reference Management

- Zotero - Reference Manager: https://www.zotero.org/
- Mendeley - Reference Manager: https://www.mendeley.com/
- EndNote - Reference Software: https://endnote.com/

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Annier Jesús Fajardo Quesada
https://orcid.org/0000-0002-2071-3716
Universidad de Ciencias Médicas de Granma. Granma, Cuba.
annierfq01@gmail.com

Eduardo Antonio Hernández González
https://orcid.org/0000-0001-7325-6099
Universidad de Ciencias Médicas de Pinar del Río. Pinar del Río, Cuba.

René Herrero Pacheco
https://orcid.org/0000-0002-9450-1572
Universidad de Ciencias Médicas de Granma. Granma, Cuba.

Lisannia Virgen Beritán Yero
https://orcid.org/0009-0004-6992-9915
Universidad de Ciencias Médicas de Pinar del Río. Pinar del Río, Cuba.

ABOUT THE AUTHOR / SOBRE EL AUTOR

Annier Jesús Fajardo Quesada was born in Jiguaní, Granma Province, on January 1, 2001. He studied at the Silberto Álvarez Aroche Pre-University Vocational Institute of Exact Sciences, where he excelled in provincial and national competitive programming competitions, winning first place at the provincial level and bronze and gold medals at the national level. He was a member of the national pre-selection team for two consecutive years and participated in regional olympiads and national cups, earning several distinctions.

Later, he entered the University of Medical Sciences of Granma, where, thanks to his research and publications, he was selected as editor of the student scientific journal Dos de Diciembre. During his university career, he participated in numerous national and international scientific events, consolidating a line of research focused on software development applied to publishing management and medical education.

He served as a professor of Mathematics and Biostatistics at the Faculty of Medical Sciences in Bayamo and as a freelance web developer. In 2025, he joined the Union of Computer Scientists of Cuba.

He was awarded a Vertical Internship in Anesthesiology and Resuscitation, where he is developing his doctoral thesis project, focused on the design of predictive indices applied to anesthetic practice.

ADDITIONAL INFORMATION / INFORMACIÓN ADICIONAL

Statement of Responsibility
The ethical and legal responsibility for the content of this work rests solely with its authors, who guarantee compliance with current regulations regarding intellectual property and copyright. The publisher is not responsible for the opinions, results, or interpretations expressed, nor for the use that third parties make of this material.

Statement of Conflict of Interest
The authors declare that they have no personal, commercial, institutional, or financial conflict of interest related to this work.

Financing
This work has not received specific funding from public, private, or non-profit organizations.